Technical Report 1151

# Army Enlisted Personnel Competency Assessment Program Phase I (Volume I): Needs Analysis

**Deirdre J. Knapp and Roy C. Campbell**
Human Resources Research Organization

October 2004

**United States Army Research Institute
for the Behavioral and Social Sciences**

**U.S. Army Research Institute**
**for the Behavioral and Social Sciences**

**A Directorate of the U.S. Army Human Resources Command**

**ZITA M. SIMUTIS**
**Director**

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical Review by

Elizabeth Brady, U.S. Army Research Institute
William Badey, U.S. Army Research Institute

## NOTICES

**DISTRIBUTION:** Primary distribution of this Technical Report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: DAPE-ARI-PO, 2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

**FINAL DISPOSITION:** This Technical Report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

**NOTE:** The findings in this Technical Report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

# REPORT DOCUMENTATION PAGE

| | | |
|---|---|---|
| 1. REPORT DATE (dd-mm-yy)<br>October 2004 | 2. REPORT TYPE<br>Final | 3. DATES COVERED (from. . . to)<br>January 6, 2003 – December 1, 2003 |

4. TITLE AND SUBTITLE

Army Enlisted Personnel Competency Assessment Program
Phase I (Volume I): Needs Analysis

5a. CONTRACT OR GRANT NUMBER
DASW01-98-D-0047/DO #45

5b. PROGRAM ELEMENT NUMBER
622785

6. AUTHOR(S)

Knapp, Deirdre J. and Campbell, Roy C.

5c. PROJECT NUMBER
A790

5d. TASK NUMBER
104

5e. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Human Resources Research Organization
66 Canal Center Plaza, Suite 400
Alexandria, VA 22314

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U. S. Army Research Institute for the Behavioral & Social Sciences
2511 Jefferson Davis Highway
Arlington, VA 22202-3926

10. MONITOR ACRONYM
ARI

11. MONITOR REPORT NUMBER
Technical Report 1151

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Contracting Officer's Representative & Subject Matter POC: Tonia S. Heffner

14. ABSTRACT *(Maximum 200 words)*:

In the early 1990s, the Department of the Army abandoned its Skill Qualification Test (SQT) program due primarily to maintenance, development, and administration costs. Cancellation of the SQT program left a void in the Army's capabilities for assessing job performance qualification. To meet this need, the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) instituted a 3-year program of feasibility research related to development of a Soldier assessment system that is both effective and affordable. The PerformM21 program has two mutually supporting tracks. The first is a needs analysis that will result in design recommendations and identification of issues related to implementation of a competency assessment program. The second track is a demonstration of concept – starting with a prototype core assessment targeted to all Soldiers eligible for promotion to Sergeant, followed by job-specific prototype assessments for several Military Occupational Specialties. Experience with the prototype assessments will influence elaboration of the operational program design recommendations.

The present report describes the needs analysis work and subsequent Army competency assessment program design recommendations as they stand at the end of the first year of the PerformM21 effort. A variety of areas are discussed, including program goals and policies as well as test content, design, development, and administration considerations.

15. SUBJECT TERMS

| | | |
|---|---|---|
| Behavioral and social science | Personnel | Job performance measurement |
| Manpower | | |

| SECURITY CLASSIFICATION OF | | | 19. LIMITATION OF ABSTRACT | 20. NUMBER OF PAGES | 21. RESPONSIBLE PERSON (Name and Telephone Number) |
|---|---|---|---|---|---|
| 16. REPORT<br>Unclassified | 17. ABSTRACT<br>Unclassified | 18. THIS PAGE<br>Unclassified | Unlimited | 82 | Ellen Kinzer<br>Technical Publications Specialist<br>(703) 602-8047 |

Technical Report 1151

# Army Enlisted Personnel Competency Assessment Program Phase I (Volume I): Needs Analysis

**Deirdre J. Knapp and Roy C. Campbell**
Human Resources Research Organization

**Selection and Assignment Research Unit**
**Michael G. Rumsey, Chief**

**U.S. Army Research Institute for the Behavioral and Social Sciences**
**2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926**

**October 2004**

**Army Project Number**
**20363007A792**

**Personnel, Performance,**
**and Training**

## FOREWORD

In April 2002, the Army Training and Leader Development Panel (ATLDP) released the results of its survey of 35,000 Noncommissioned Officers (NCOs). The ATLDP's recommendations included the need for regular assessment of Soldiers' technical, tactical, and leadership skills. The need for regular assessment of Soldiers coincides with the U.S. Army Research Institute for the Behavioral and Social Sciences' (ARI) research program on NCO development and assessment. ARI's research program began with *Soldier Characteristics of the 21st Century (Soldier21)* to identify potential knowledges, skills, and attributes (KSAs) for future Soldiers and continued with *Maximizing 21st Century Noncommissioned Officers Performance (NCO21)* to identify and validate potential indicators of the KSAs for use in junior NCO promotion. The *Performance Measures for 21st Century Soldier Assessment (PerformM21)* extends the research program with a three-phase effort to examine the feasibility of comprehensive competency assessment. The first phase is an investigation of the issues and possible resolutions for development of a viable Army-wide program including the Demonstration Competency Assessment Program (DCAP), which is a prototype for Army-wide competency assessment. The second phase extends the feasibility investigation through development of five Military Occupational Specialties (MOS) competency assessments as well as a self-assessment and development module to accompany the DCAP. The third phase is an analysis of the prototype program to provide recommendations on feasibility, resource requirements, and implementation strategies for competency assessment. This multi-volume report documents activities supporting the first goal of Phase I—issues impacting overall recommendations for Army-wide assessment—and also describes the development of the DCAP assessment. The prototype DCAP assessment and elements of the recommended delivery system will be pilot tested in Phase II of the project. Program design issues identified here will inform future deliberations about the design, implementation, and maintenance of an operational assessment program.

The research presented in this report has been briefed to the Deputy Chief of Staff, G-1, on 8 Oct 2003 and the Chief of Enlisted Professional Development, Directorate of Military Personnel Policy on 13 Nov 2003. It was briefed to the Sergeant Major of the Army on 28 Jan 2003 and 30 Mar 2004. It has been periodically briefed to senior NCO representatives from U.S. Army Training and Doctrine Command (TRADOC), Office of the G-1, U.S. Forces Command (FORSCOM), U.S. Army Reserve (USAR), and the Army National Guard (ARNG) as members of the Army Testing Program Advisory Team (ATPAT).

The goal of ARI's Selection and Assignment Research Unit is to conduct research, studies, and analysis on the measurement of attributes and performance of individuals to improve the Army's selection and classification, promotion, and reassignment of officers and enlisted Soldiers.

Paul a. Gade

PAUL A. GADE
Acting Technical Director

# Acknowledgements

## U.S. Army Research Institute for the Behavioral and Social Sciences (ARI)<br>Contracting Officer Representatives (COR)

Dr. Peter Greenston and Dr. Tonia Heffner of ARI served as co-COR for this project, but their involvement and participation went far beyond the usual COR requirements. Their contributions and active input played a significant role in the production of the final product and they share credit for much of the outcome. Of particular note are their activities in conveying information about the project in briefings and presentations to Army Leadership on many important levels.

## The Army Test Program Advisory Team (ATPAT)

The functions and contributions of the ATPAT, as a group, are documented in this report. But this does not fully reflect the individual efforts that were put forth by members of this group. Project staff is particularly indebted to CSM Cynthia Pritchett, Command Sergeant Major, U.S. Army Combined Arms Center and Fort Leavenworth. CSM Pritchett not only serves as the ATPAT Chairperson but has provided wise counsel and guidance in a number of distinct areas since the inception of the project in January 2003. It was through her initiative and recommendations that the ATPAT was established.

Serving as co-chair of the ATPAT is SGM Michael T. Lamb, Sergeant Major, U.S. Army Training and Doctrine Command, Deputy Chief of Staff for Operations and Training, Fort Monroe, Virginia. His involvement also, has transcended the ATPAT activities described and we have come to rely on his assistance and involvement in areas too numerous to detail.

The other individual members of the ATPAT are:

SGM Thomas Clark
U.S. Army Human Resources Command

SGM Fredrick Couch
ARNG Training Division Sergeant Major

SGM John T. Cross
Center for Army Leadership (CAL)

CSM George D. DeSario
Command Sergeant Major
U.S. Army Armor Center

SGM Julian Edmondson
Personnel Policy Integrator, Army G-1

CSM Dan Elder
Command Sergeant Major
HQ, 13th Corps Support Command

CSM Victor Gomez
Command Sergeant Major
HQ, 95th Division

SGM Paul Harzbecker
G-3 Sergeant Major
HQ, Forces Command (FORSCOM)

SGM James Herrell
Sergeant Major, Ordnance Proponency

SGM Enrique Hoyos
Sergeant Major, Army Training Support
Center (ATSC), TRADOC

CSM Nick Piacentini
Command Sergeant Major
U.S. Army Reserve Command (USARC)

SGM Gerald Purcell
Directorate Sergeant Major
Military Personnel Policy, Army G-1

CSM Robie Roberson
Group Command Sergeant Major
653rd Area Support Group

CSM Otis Smith Jr.
Command Sergeant Major
U.S. Army Armor School

CSM Clifford R. West
Command Sergeant Major
U.S. Army Sergeants Major Academy

MSG Daphne Angell
309th Regiment, 78th Division

MSG Robert Bartholomew
Enlisted Career Manager
Ordnance Proponency

MSG Monique Ford
Operations NCOIC
309th Regiment, 78th Division

MSG Fred Liggett
Promotion Policy Integrator, Army G-1

MSG Christopher Miele
Operations NCOIC
14th The Army School System (TASS)
Battalion

MSG Matt Northen
G-3, Forces Command (FORSCOM)

1SG Edwin Padilla
First Sergeant, HQ and HQ Detachment
2nd Battalion, 309th Regiment, 78th Division

MSG Jerome Skeim
Enlisted Manager for Reclassification
National Guard Bureau

ARMY ENLISTED PERSONNEL COMPETENCY ASSESSMENT PROGRAM PHASE I (VOLUME I): NEEDS ANALYSIS

## Executive Summary

Research Requirement:

The Army Training and Leader Development Panel NCO survey (Department of the Army, 2002) called for objective performance assessment and self-assessment of Soldier technical and leadership skills to meet emerging and divergent Future Force requirements. The Department of the Army's previous experiences with job skill assessments in the form of Skill Qualification Tests (SQT) and Skill Development Tests (SDT) were effective from a measurement aspect but were burdened with excessive manpower and financial resource requirements.

Procedure:

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) is conducting a 3-year feasibility effort to identify viable approaches for the development of a useful yet affordable operational performance assessment system for Army enlisted personnel. Such a system would depend on technological advances in analysis, test development, and test administration that were unavailable in the previous SQT/SDT incarnations.

The ARI project (known as PerformM21) is being conducted with support from the Human Resources Research Organization (HumRRO) and entails three phases:

- Phase I: Identify User Requirements, Feasibility Issues, and Alternative Designs
- Phase II: Develop and Pilot Test Prototype Measures
- Phase III: Evaluate Performance Measures and Make System Recommendations

The objective of Phase I was to isolate and identify issues that the overall recommendation needs to take into account for a viable, Army-wide system. This is the topic of the present report. Phase I also produced a rapid prototype assessment covering Army-wide "core content" in the form of a Demonstration Competency Assessment Program (DCAP), which is documented in a separate report (Campbell, Keenan, Moriarty, Knapp, & Heffner 2004).

In Phase II, the ARI/HumRRO research team will (a) pilot test the DCAP, (b) develop competency assessments for up to five Military Occupational Specialties (MOS), and (c) explore issues further to develop more detailed recommendations related to the design and feasibility of a new Army enlisted personnel competency assessment program. In Phase III, the concept and pilot testing of the assessment program will be evaluated.

Findings:

The envisioned competency assessment program would be implemented in phases, starting with an Army-wide core assessment suitable for enlisted Soldiers at the

specialist/corporal level (E4). The demonstration core assessment (i.e., the DCAP) includes sections on common Soldiering skills, leadership, training, values, and history. The program would then expand to include core assessments for higher enlisted grade levels (E5 through E7) and to include MOS-specific assessments. To the extent possible, assessments would be delivered through Army Digital Training Facilities (DTFs) in annual test windows (e.g., 60-90 day periods) thus allowing sufficient opportunity for Soldiers in a variety of settings (e.g., Reserve unit training weekends, deployments) to participate. Some MOS-specific assessments, however, could require supplemental delivery models to accommodate certain desired assessment methods (e.g., hands-on testing).

In addition to discussing details associated with the various aspects of the assessment program (e.g., test design, development, maintenance, and delivery), several overarching considerations are relevant. These include the need for the Department of the Army to (a) prepare to bear the cost of the program, (b) carefully consider the timing of implementation, (c) obtain buy-in from stakeholders, and (d) commit to both cost-effectiveness and quality.

Utilization of Findings:

The ideas and issues identified in Phase I of the PerformM21 research program will be further explored, defined, and evaluated in subsequent phases of the effort. The end result will be increasingly fine-tuned recommendations and prototype assessments and procedures that will aid the Army in building a new, effective competency assessment program for selecting and growing NCOs.

ARMY ENLISTED PERSONNEL COMPETENCY ASSESSMENT PROGRAM PHASE I (VOLUME I): NEEDS ANALYSIS

## CONTENTS

## List of Appendices

## List of Tables

## List of Figures

# ARMY ENLISTED PERSONNEL COMPETENCY ASSESSMENT PROGRAM PHASE I (VOLUME I): NEEDS ANALYSIS

## Chapter 1. Introduction

### *Background*

The Department of the Army is changing to meet the needs of the 21st century. Soldiers at all levels must possess the interpersonal, technical, and organizational knowledge, skills, and other attributes to perform effectively in complex technical, information-rich environments, under multiple and changing mission requirements, and in semi-autonomous, widely dispersed teams. The Army needs an integrated Soldier assessment system to support these demands.

The need for Soldier assessment is most acute at the time of promotion in the Noncommissioned Officer (NCO) ranks. It is at this juncture that job competency merges with leadership and supervisory requirements and there are distinct changes in the concept of Soldiering. In June 2000, the Chief of Staff of the Army established the Army Training and Leader Development Panel (ATLDP) to chart the future needs and requirements of the NCO corps. After a 2-year study which incorporated the input of 35,000 NCOs and leaders, a major conclusion and recommendation was: "Develop and sustain a competency assessment program for evaluating Soldiers' technical and tactical proficiency in the military occupational specialty (MOS) and leadership skills for their rank" (Department of the Army, 2002).

In the early 1990s, the Army abandoned its Skill Qualification Test (SQT) program due primarily to maintenance, development, and administration costs. Cancellation of the SQT program left a void in the Army's capabilities for assessing job performance qualification. Re-instituting a new performance assessment system must address the factors that forced abandonment of the SQT. Since then, technological advances have occurred that can reduce the developmental and administrative burdens encountered with SQT and will play a critical role in a new performance assessment system.

To meet the Army's need for job-based measures, the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) instituted a 3-year program of feasibility research to identify viable approaches for development of a Soldier assessment system that is both effective and affordable. This research is being conducted with contract support from the Human Resources Research Organization (HumRRO). The impetus to include individual Soldier assessment research in ARI's programmed requirements began prior to 2000 and was based on a number of considerations regarding trends and requirements in Soldier selection, classification, and qualifications. Meanwhile, there were several significant events within the Army that coincided with ARI's efforts in this area. The aforementioned ATLDP recommendation resulted in the Office of the Sergeant Major of the Army (SMA) and the U.S. Army Training and Doctrine Command (TRADOC) initiating a series of reviews and consensus meetings with the purpose of instituting a Soldier competency assessment test. Ongoing efforts within the Army G-1 to revise the semi-centralized promotion system (which promotes Soldiers to the grades of E5 and E6) also were investigating the use of performance (test)-based measures to supplement the administrative criteria used to determine promotion. Ultimately, the three interests (ARI, SMA/TRADOC, G-1) coalesced and the ARI project sought to incorporate the program goals and operational concerns of all of Army stakeholders, while still operating within its research-mandated orientation.

*PerformM21*

The ARI program (called PerformM21) has three phases:

- Phase I: Identify User Requirements, Feasibility Issues, and Alternative Designs
- Phase II: Develop and Pilot Test Prototype Measures
- Phase III: Evaluate Performance Measures and Make System Recommendations

Phase I of the program (which corresponds roughly to year one of the 3-year overall effort) had three primary goals:

- Goal 1: Determine the feasibility and inherent trade-offs in the development of an operational and affordable individual performance assessment system for Army enlisted Soldiers.
- Goal 2: Identify the major design considerations and elements of such a system.
- Goal 3: Develop a prototype assessment measure.

The research program is best viewed as having two mutually supporting tracks. The first track (reflected in Goals 1 and 2 above) is essentially a needs analysis that results in design recommendations and identification of issues related to implementation of a new enlisted Soldier competency assessment program. The second track is a rapid prototype demonstration of concept – called the Demonstration Competency Assessment Program (DCAP). The DCAP is intended to reflect, inasmuch as possible, design recommendations for the future operational assessment program. Experience with the DCAP will in turn influence elaboration or modification of the operational program design recommendations as they develop during the course of the 3-year research program.

The DCAP assessment covers Army-wide "core content" in the form of a computer-administered, largely multiple-choice examination. Development of the DCAP is documented in a separate technical report (Campbell, Keenan, Moriarty, Knapp, & Heffner, 2004). It will be pilot tested in 2004. Also in 2004, several prototype assessments covering competencies specific to selected Military Occupational Specialties (MOS) will be developed. These assessments will demonstrate a broader array of assessment methods (e.g., computer simulations, hands-on tests). They will be pilot tested and final recommendations made as part of the Phase III evaluation scheduled for 2005.

*The Army Test Program Advisory Team (ATPAT)*

Early in Phase I, we constituted a group to advise on the operational implications of Army assessment testing, primarily as part of the needs analysis aspect of the project.

Simultaneously, this group took on a role as Test Council for the DCAP. This group is called the Army Test Program Advisory Team (ATPAT) and it has the following characteristics:

- It is made up of NCOs, mostly in the Master Sergeant (E8) and Sergeant Major (E9) levels.
- It includes representatives from TRADOC, HQ, Forces Command (FORSCOM), Combined Arms Center (CAC), Center for Army Leadership (CAL), Army Training Support Center (ATSC), Army G-1, Sergeant Major Academy (USASMA), and specific organizational representation including the U.S. Army Armor Center and School, the U.S. Army Ordnance Center and School, and the 13 Corps Support Command (COSCOM).
- It includes representatives from the Reserve force including from HQ, Army National Guard Bureau (ARNGB), HQ, Army Reserve Command (USARC), and unit representatives from the 95th Division (Institutional Training), 653rd Area Support Group (ASG), and the 78th Division (Training Support).
- It is co-chaired by two Sergeants Major endorsed by the ATPAT. Chairs are responsible for approving agenda items for meetings and for serving as points of contact for the ARI/HumRRO project team.
- It has a flexible membership. Although there is a solid core ATPAT group, there have thus far been 25 individual representatives to the ATPAT.

The ATPAT serves two distinct purposes. First, it provides the primary input for the needs analysis requirements of the project, primarily by providing insight into operational implications and real-world feasibility of the program. Second, it serves as the oversight group for development of the DCAP as well as a resource in identifying and developing content for the test. Additionally, the ATPAT is a working group that provides product reviews, subject matter expertise, and, as needed, assistance in the process of developing prototype instruments and trial procedures. An additional benefit of the ATPAT is to serve as a conduit to explain and promote the PerformM21 project to various Army agencies and constituencies.

The ATPAT met three times in 2003, providing guidance in four areas:

- *Utilization Strategies* – How the test will be used, defining and limiting the scope of the program. That is, whether and how it will be used in personnel management, promotion, career development, training, readiness, retention, and transition.
- *Implementation Strategies* – Identifying the steps to implementing, maintaining, and growing an Army test, short- and long-term goals, and organizational implications to be considered in phased implementation.
- *Operational Strategies* – Identifying the considerations that must be taken into account to operationalize an Army-wide testing program (for developers, administrators, and users).

- *External Considerations* – How the Army-wide test will fit in with other agendas such as self-development, unit training, the NCO Education System (NCOES), deployments, the NCO Evaluation Report (NCOER), Table of Distribution Allowances (TDA) staffing, transition, Soldier tracking and assignment, Future Force, and training publications and updates.

The ATPAT has been extremely helpful in discussions in all of these areas. At each meeting, significant portions of the discussion were centered on the nature of the assessment (i.e., self-assessment vs. promotion). Other significant discussion and activities centered on the determination and specification of the content domains of the DCAP. They have also helped to identify resources useful for test item development and field testing.

*Needs Analysis Organizing Structure and Process*

To structure the needs analysis process, project staff drafted a list of requirements for supporting an assessment program. Figure 1 lists the key components, with further detail about each component provided in Appendix A. This structure helped organize our thinking and suggested the questions we posed to those providing input into the process. We obtained input from several sources as we considered the issues, ideas, and constraints associated with each requirement listed in Figure 1. These included the following:

- The Army Testing Program Advisory Panel (ATPAT)
- Historical information about the SQT program and associated lessons learned
- Enlisted personnel promotion testing programs operated by the Air Force and the Navy
- Civilian assessment programs (e.g., professional certification programs)
- A review of automation and technology tools and systems
- Work completed under a related project

The remainder of this section briefly describes these resources.

*Army Testing Program Advisory Panel (ATPAT).* As described previously, the ATPAT was an important source of input into the needs analysis process. This group of Army experts reviewed the organizing structure outlined in Figure 1 for completeness and provided ideas and reactions to proposals associated with the various requirements. Most of the recommendations described in the remainder of this report came directly from this group.

*Historical information about SQT.* HumRRO project staff members were closely involved with development of the former SQT program and are knowledgeable about what worked well and what did not. Appendix B provides an historical review that summarizes what we know about the SQT experience. To the extent possible, lessons learned from this experience are being integrated into the PerformM21 needs analysis recommendations.

*The other services.* At the onset of Phase I, project staff visited the U.S. Air Force and U.S. Navy offices responsible for administering their respective enlisted personnel promotion testing programs. Both services have extensive experience in assessment testing. Summaries of what we learned about these programs are provided in Appendixes C and D. Neither program provides a model that is exactly right for the Army. For example, at this point both programs are paper-based, although the Navy is planning for eventual transition to computer-based test delivery. There are, however, important lessons to be learned from these models.

- Purpose/goals of the testing program
- Test content
- Test design
- Test development
- Test administration
- Interfacing with candidates
- Associated policies
- Links to Army systems
- Self-assessment

*Figure 1. Outline of PerformM21 needs analysis organizing structure.*

*Civilian testing programs.* Another place to look for ideas and lessons learned is the civilian world. Large-scale competency assessment programs are evident primarily in the arena of professional certification. This is a growing industry that has expanded from a concentration in healthcare occupations (e.g., physician and nursing specialty certifications) to include a wide array of technical and managerial jobs. HumRRO is active in the civilian credentialing community and brings that perspective to thinking in the needs analysis. As the research program progresses, it will be useful to develop systematic strategies for collecting information and potential partnering relationships between the Army and civilian programs.

*Automation and technology.* We organized a team of HumRRO personnel with particular knowledge and experience of technology as it applies to large-scale assessment to learn about the Army's existing capabilities and craft recommendations related to this area. We call this team the Core Technology Group and will keep it as a resource throughout the project. This is an important area because of our a priori assumption that advances in technology will be key to designing an assessment program that will be feasible and affordable. Appendix E summarizes the recommendations this group is making with regard to test development and delivery software, delivery portal, and web-delivery support requirements.

*Related project.* In February through August 2003, Job Performance Systems, Inc. (JPS) and HumRRO (subcontracting to JPS) worked on a grant that was designed to complement the PerformM21 research program (Rosenthal, Sager, & Knapp, 2003). The primary product of this effort was a proposed methodology for the Army to use to produce realistic and cost-effective job performance assessments. We will incorporate elements of that proposed methodology into the PerformM21 project plan. For example, that methodology will help guide our approach to the job analysis work that will be required to support MOS test development.

*Overview of Report*

The remainder of this report discusses the needs analysis results, in terms of recommendations and issues, as they stand at the conclusion of Phase I. Specifically, Chapter 2 provides an overall vision of the new Army assessment program as it is emerging, given input from the sources listed above. The remaining chapters provide more detail about key aspects of the program. These include:

- Test specifications
- Test development
- Test delivery
- Interfacing with Soldiers
- Integrating the assessment program into Army systems

As we move into Phase II of the PerformM21 research program, the vision of the new program will no doubt evolve and become more detailed. In the meantime, this report is intended to be a snapshot that captures where the concept stands at this moment in time.

# Chapter 2. An Overall Vision of the Soldier Assessment Program

The design, implementation, and maintenance of an effective and cost-efficient competency assessment program are daunting challenges. On the other hand, the payoffs of such a program are both far-reaching and critical to the continued effectiveness of the U.S. Army. Readiness begins with the capabilities of individual Soldiers. Effective Soldier selection and classification followed by high quality career-long training and development programs are the primary tools for obtaining superior Soldiers. Periodic assessment of Soldier knowledge and skills is required to provide a scorecard that will help ensure accountability of individual Soldiers and the Army systems that are supposed to support Soldiers' training and development needs.

## *The Program*

Figure 2 lists the basic features of the envisioned competency assessment program. The program would be implemented in phases, starting with an Army-wide core assessment suitable for enlisted Soldiers at the Specialist/Corporal level (E4). The demonstration core assessment (i.e., the DCAP) includes sections on common Soldiering skills, leadership, training, values, and history. The program would then expand to include core assessments for higher enlisted grade levels (E5 through E7) and to include MOS-specific assessments. To the extent possible, assessments would be delivered through Army Digital Training Facilities (DTFs) in annual testing windows (e.g., 60-90 days), thus allowing sufficient opportunity for Soldiers in a variety of settings (e.g., Reserve unit training weekends, deployments) to participate. Some MOS-specific assessments, however, are likely to require supplemental delivery models to accommodate certain assessment methods that may be desired (e.g., hands-on testing).

While it is understood that the assessment program must yield performance indicators that can be incorporated into the enlisted promotion system (i.e., through inclusion on a revised Promotion Point Worksheet), it is important that Soldiers be supported by the program inasmuch as possible. This will be accomplished in at least two ways. First, test preparation guides will be available to help Soldiers prepare for the assessment. Second, Soldiers will be given feedback on their performance so they have a better sense of their relative strengths and weaknesses as indicated by the assessment. We also recommend the first Army-wide administration of the examination be a dry run, with scores not used for promotion points until the second time around.

Although outside the scope of the assessment program, per se, a self-assessment program is under development. The initial version will be targeted to knowledge and skills covered by the core assessment. In the self-assessment exercise, Soldiers will get feedback on their responses to individual test questions and scores will not become part of their official records. The proposed operational self-assessment program would have a broader focus – helping Soldiers gauge their promotion potential with regard to all elements included on the Promotion Point Worksheet. To effectively serve Soldiers' needs, both the test preparation guides and self-assessment exercises must be backed up with up-to-date training material and job aids.

Core assessment + MOS assessment

- Initially core assessment only
- MOS assessments would be added as they become available
- Core and MOS assessments would likely be administered separately
- MOS assessments should reflect the distinct characteristics of each job; they could vary widely in their focus and content

Assessment method

- Core assessment will include machine scoreable items (e.g., multiple-choice, drag and drop, matching) designed to be realistic rather than textbook-like
- MOS assessments may include other assessment methods, such as hands-on tests and high fidelity computer-based simulations

Annual testing

- Start with an annual testing window (e.g., 60-90 days) for the core assessment
- As MOS assessments become available, add testing windows that would spread testing throughout the year (e.g., Career Management Field (CMF) 91 tests in Jan-Feb; CMF 63 tests in Mar-Apr, and so forth)

Assessment delivery

- Most assessments scheduled through and administered on-line at Army Digital Training Facilities (DTFs)
- Some MOS-specific assessments may require different delivery mechanisms (e.g., for hands-on assessments or high fidelity simulations)

Scoring

- Test results reviewed and finalized by the test authority (either a new agency or an existing ATSC organization, likely with supporting contractors) at conclusion of administration period
- Scores entered into Soldier records and applicable promotion points allocated
- Feedback reports emailed to Soldiers

Soldier preparation

- Test preparation guides available on the Army Knowledge Online (AKO) website at least 120 days before assessments are administered; preparation guides will not significantly differ from one year to the next
- Self-assessment exercise will be available to all before core assessment is given for scores of record
- Test preparation guides and the self-assessment program will refer Soldiers to training and job aids (e.g., Soldier manuals) that need to be up-to-date and available for their use

*Figure 2. Overview of assessment program major design features.*

*Supporting Structure and Functions*

### Oversight and Coordination

Figure 3 illustrates some of the support systems that the Department of the Army will need to accommodate its new assessment program. First, the Army should ideally identify or establish individuals and offices to be responsible for policy decisions, oversight, and coordination of the assessment program across all components (Active and Reserves). Unlike the former SQT program, the new program will have a core assessment that does not have an existing single proponent. Moreover, it will be necessary for a cost-effective program to have coordination among MOS proponents that exceeds that required in the old SQT program. Such a group would meet periodically to determine strategies for sharing resources and lessons learned. The group would also help ensure that differences in MOS assessment policies and procedures make sense in light of MOS differences rather than being haphazard in nature across the Army.

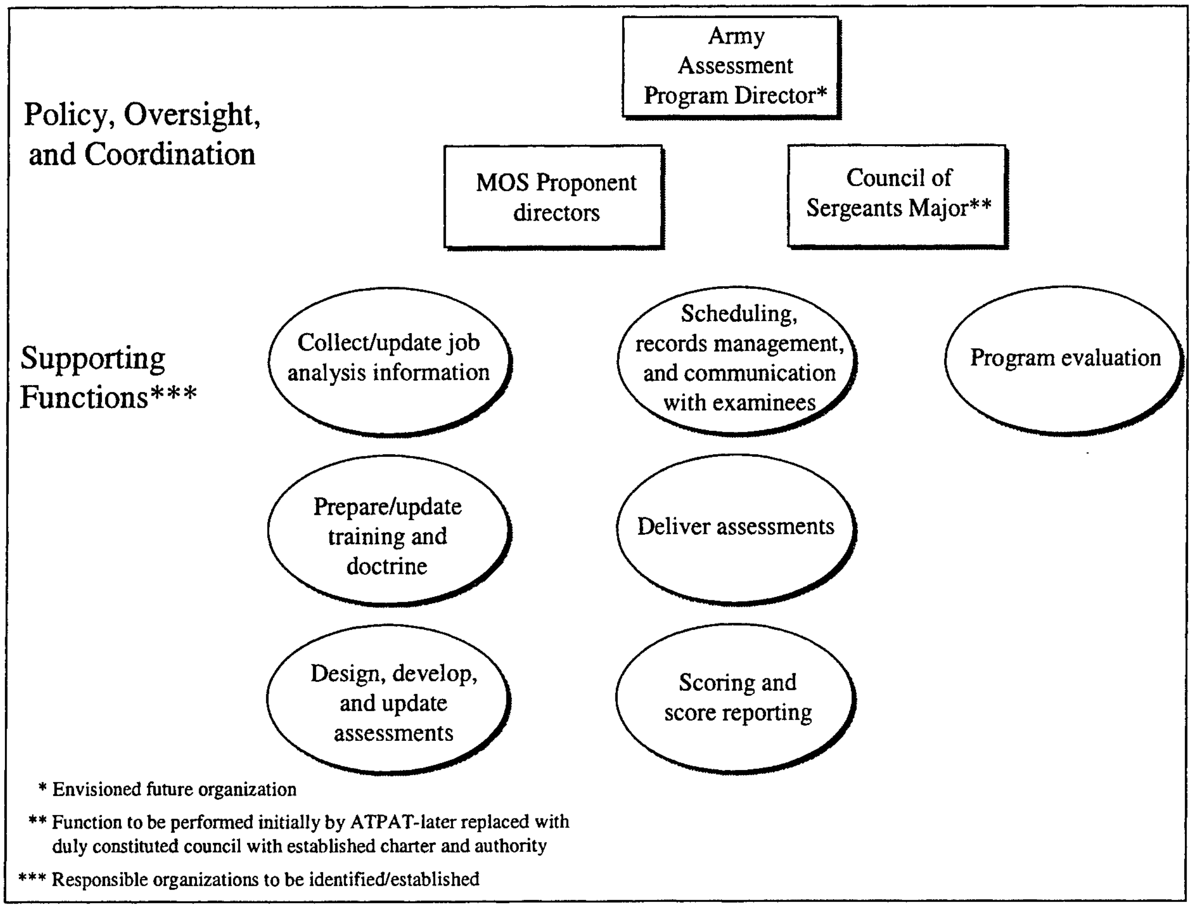
Army Assessment Program Director*
Policy, Oversight, and Coordination
MOS Proponent directors
Council of Sergeants Major**
Supporting Functions***
Collect/update job analysis information
Scheduling, records management, and communication with examinees
Program evaluation
Prepare/update training and doctrine
Deliver assessments
Design, develop, and update assessments
Scoring and score reporting
* Envisioned future organization
** Function to be performed initially by ATPAT-later replaced with duly constituted council with established charter and authority
*** Responsible organizations to be identified/established

*Figure 3. Assessment program supporting structure and functions.*

The Army Assessment Program Director would be the officer responsible for overall Army test program policy in development and administration. This office would also be responsible for overseeing and coordinating policy within the various MOS proponents and for coordinating test use and implementation among the many Army agencies and commands. (See

Chapter 7 for a more detailed discussion of organization considerations for this office.) The Director should be a senior officer at the Colonel (O6) level.

A Council of Sergeants Major would be responsible for making recommendations related to the design, content, and policies associated with the core assessment program. In the development of the DCAP, this function has been served by the ATPAT. It is reasonable to think the ATPAT could evolve into the standing Council of SGMs with suitable guidelines for membership, composition, and responsibilities.

### Supporting Functions

The lower part of Figure 3 notes the major functions that are required to support the envisioned assessment program. They are loosely organized into what is needed to (a) develop and maintain the assessments (i.e., job analysis data; training and job aids on which to base test content; people and systems to design, develop, and update the assessments), (b) administer the assessment program (e.g., scheduling and delivering the assessments, scoring and score reporting), and (c) evaluate the program. None of these functions is optional. For example, program evaluation is necessary early on to ensure problems are identified and corrected quickly. It is also necessary in the long term, to help ensure that the assessment program evolves to optimize application of new technologies and lessons learned in a system of continuous process improvement.

Some of these functions are already in place (e.g., collecting/updating job analysis information, preparing/updating training and doctrine), but they are not currently adequate to support the needs of the assessment program. To be fair to Soldiers and effectively serve the Army's needs, the assessments must reflect current job requirements and Soldiers must have access to training materials (e.g., technical manuals, Soldiers' manuals) that are accurate, complete, and up-to-date. For a variety of reasons, this is not uniformly the case. Moreover, the type of job analysis information needed to support development of high quality assessments is somewhat different from that currently collected through the Army's occupational analysis program (see Army Regulation 370-50). It makes sense, however, to modify current job analysis procedures to support the new program's needs rather than proposing an entirely new or largely redundant system.

### *Guiding Principles*

The following chapters discuss in detail our current thinking and recommendations about various aspects of the envisioned assessment program outlined in this chapter. Before moving to that level of detail, however, there are some general conclusions that we have reached that should be kept in mind as guideposts and reminders throughout:

- *The Art of Assessment:* We know a great deal about how to develop good assessments – ones that are valid and psychometrically reliable and that make optimal use of advanced technology. What presents a particular challenge for the present situation is designing an assessment program that will be high quality, cost-effective, and practically supportable by the Army. If resources were not an issue, this would be relatively easy.

- *The Cost of Assessment:* As should be clear already and will become even more evident as discussions and planning continue, an effective assessment program that includes core and MOS-specific assessments for Soldiers at multiple pay grades will be an expensive undertaking. Generally speaking, however, start-up costs will be considerably higher than maintenance costs. One advantage to rolling out the assessment program in phases is to make the upfront costs more manageable.

- *The Timing of the Assessment Program:* There are other significant advantages to growing and developing the assessment program over time. First, the program must develop the trust of Soldiers. While it is true that examinees will always criticize the tests they have to take, the Army can accomplish a great deal by making sure that Soldiers learn to respect this new assessment program. Early problems can damage the program's reputation in ways that are virtually impossible to overcome once that damage is done. Our advice is to start slow and strong and manage expectations as much as possible. This means being clear that the program will grow and improve with experience and input from all stakeholders.

- *Buy-In to Assessment*: It is not just individual Soldiers who need to be convinced the assessment program is worth doing and doing well. The more that all supporting organizations and individuals are vested in and respect the idea, the more effective it will be. It will therefore be important to "market" the program to all stakeholders by informing them of program goals and plans while also educating them about the various considerations and constraints that go into developing a successful program. People invariably underestimate what it takes to develop and maintain a high quality assessment program and they often question the motives of those developing those programs.

- *Quality of the Assessment Program:* Finally, if the Army is serious about instituting an assessment program such as that outlined here, it must be willing to support it without cutting corners. For as important as it is to have, a low quality assessment program would be worse than none at all. A poorly maintained program would be unfair to Soldiers and would not help the Army improve readiness. Commitment to quality is critical to the program's success.

# Chapter 3. Test Specifications

This chapter describes a process for deciding what content should be included in the Army assessments. The details will evolve, but the basic process is that used in the development of certification, licensure, employee selection, and other assessments used throughout U.S. government, business, and industry. As we will do in the chapters to follow, we begin with a brief discussion of some of the challenges the Army faces with regard to this step in the process.

## *Challenges*

Up-to-date, detailed occupational analysis (i.e., job requirement) information is fundamental to determining the appropriate content for competency assessments. All of the U.S. Services, and some NATO allies, maintain occupational survey and analysis programs. The Army has an occupational analysis system that involves the periodic collection of data on job tasks and associated knowledge and skills (see Army Regulation 370-50). This is known as the Occupational Analysis Survey Program (AOSP) (see AR 611-3). The ARI Occupational Analysis Office supports Department of the Army (DA) personnel proponents and schools involved in Manpower, Personnel, and Training studies. This includes providing TRADOC occupational survey development, data collection, and analysis support in accordance with TRADOC Regulation (TR) 350-70 (Systems Approach to Training Management, Processes, and Products) and DA Regulation 350-1 (Army Education and Training).

The current AOSP information is insufficient to support the new assessment program for several reasons:

- Occupational analysis information for many MOS is either unavailable or out-of-date (i.e., more than 3 years old).
- The occupational analysis information is of uneven format and quality (e.g., with regard to task detail, clarity, comprehensiveness, and overlap); adequacy of Soldier samples is also a concern, particularly when considering today's high level of deployment activity.
- For assessment purposes, it may be desirable to focus on less detailed job tasks and to use different strategies to identify relevant knowledge and skills than specified by TR 370-70.
- Certain assessment methods that are likely to be adopted will require types of information not currently provided (e.g., critical incidents, walk-through analysis)

The first two problems listed above are largely resource issues dating back to the mid-1990s when the AOSP experienced a severe depletion in resources. In addition, training products (e.g., course curricula, Soldiers' manuals) – the intended beneficiary of the AOSP – are suffering as a result (General Accounting Office, 2003). Ideally, we could recommend a system that would build on the AOSP so that the Army does not have to embark on a totally new or largely redundant system.

More so than at any other time in its history, the Army is experiencing a rapidly evolving force structure (in terms of the number and composition of MOS). The current trend toward

MOS consolidation simplifies matters on one hand because it reduces the number of MOS-specific assessment programs that will be required. On the other hand, it increases the degree to which there are wide variations in what Soldiers with the same MOS designation actually do on their jobs. This complicates efforts to design assessments that accurately correspond to Soldiers' job requirements. It is also increasingly difficult to keep up with changes in job requirements that are much more rapid than they were during the days of the SQT program.

*Available Resources*

Fortunately, the Army also is positioned well to address the need for improved job analysis information. First, there is precedent for systematically collecting such data through the Army's Occupational Data Analysis, Requirements, and Structure (ODARS) Program that has been in place (in one form or another) for several decades. Improving and building on an existing system has many advantages over starting from scratch. Second, the civilian world has seen an explosion in occupational certification programs that cover jobs that may overlap at least in part with various Army MOS. Not all certification programs are created equal – some are based on high quality occupational information and others are not. High quality civilian certification programs present a potential source of information that might be of use to the Army. This might include occupational analysis information (that could be used to feed into development of Army occupational analyses and to compare/contrast with related Army MOS) and self-development tools (e.g., distance learning programs). Because credentialing programs do not typically provide score information (but rather just pass/fail), however, it is unlikely they could completely replace the requirement for MOS-specific competency testing.

*Job Analysis*

Rosenthal et al. (2003) proposed a methodology to facilitate the process of determining what job analysis information is required given the most likely assessment methods that will be used for a given MOS. The methodology makes use of a tentative clustering of MOS based on likely applicable assessment methods. For example, the largest cluster includes MOS for which the Army would need to develop multiple-choice exams and at least one indicator of job proficiency (e.g., computer simulations, hands-on tests, evaluation of final work products, embedded on-the-job monitoring) to provide comprehensive and realistic assessment of MOS-specific competence.

Job analysis in the Rosenthal et al. model begins with a preliminary stage in which information is collected from existing resources and a small number of subject matter experts (SMEs). This information is used to finalize decisions on assessment methods and thus determines what information needs to be generated in the full-scale job analysis. Depending on the results of the preliminary job analysis, the full job analysis will include some combination of the following strategies, most of which are incorporated into the ODARS Program:

- Direct observation of job incumbents performing their work
- Analysis of available documents/information (e.g., Soldier manuals)
- SME interviews (one-on-one)
- SME workshops

- Job analysis surveys
- Critical incidents/scenarios

Once the job analysis process is completed the first time for an MOS in a manner that supports competency testing, it should be much easier to periodically update the information as needed. It is likely that we will generate recommendations for determining when new occupational analysis information should be collected to help ensure assessments stay current with field requirements.

The Army has traditionally defined job requirements in terms of very specific job tasks. To be successful in the long run, however, the tests used in an annual competency assessment program should be based on somewhat more broadly defined requirements. This will be invaluable in helping to ensure that test specifications do not shift significantly from one year to the next (which makes it difficult for both Soldiers preparing for the tests and test developers who are creating the tests) and for accommodating differences in the specific tasks performed by Soldiers in the same MOS in different locations/units. Looking at more broadly defined tasks and the knowledge and skills required to perform those tasks (e.g., knowledge of certain underlying principles) will also make it easier to assess Soldiers doing similar activities with different pieces of equipment.

### *Assessment Methods*

Although the DCAP only uses objectively scored questions (e.g., multiple-choice, drag and drop, matching), other assessment methods can provide additional, important information about job competence. Alternative assessment methods are listed in Table 1. Typically, the most fair and useful assessments include multiple methods to triangulate on the question of whether an individual has the required job knowledge and skill.

### *Test Specifications*

As the Army's new assessment program gets off the ground, it will be important to design an assessment process for each MOS (and pay grade within the MOS) that accurately and fairly reflects job requirements. The job analysis identifies the job requirements and the alternative assessment methods provide a menu of options. The test specifications provide a recipe for the assessment. The test specifications should indicate what types of assessments will be developed and exactly what they will look like. For a multiple-choice exam, for example, the test specifications would include a blueprint that specifies (a) how many items will be on the test, (b) what content areas will be covered on the test, and (c) how many items will be in each content area.

For the DCAP core examination, there was insufficient time to collect detailed job analysis information or to seriously consider assessment methods other than multiple-choice. Thus, the test blueprint we developed with input from the ATPAT is not adequate for the operational program. Moreover, the DCAP blueprint is largely task-based. As noted previously, we recommend the Army adopt a somewhat less detailed focus for specifying test content, since detailed tasks change frequently. Having test content specifications that change quickly will be frustrating for Soldiers and make the testing program much harder to maintain.

*Table 1. Potential Assessment Methods*

| | Method |
|---|---|
| 1 | On-the-job monitoring/assessment (embedded in relevant systems) |
| 2 | Expert evaluation of actual work products |
| 3 | High fidelity hands-on work sample tests (process and/or product scoring) that closely model task requirements |
| 4 | Lower fidelity hands-on work sample tests (process and/or product scoring) requiring skipping or talking through aspects of performance (e.g., due to time/safety considerations) |
| 5 | High fidelity computer-based simulation (programming, equipment, and props to closely match the real thing) that may make use of Army training simulators |
| 6 | Medium fidelity computer-based simulation (programming, visuals, and low cost PC attachments to obtain a reasonable approximation of the real thing) |
| 7 | Low fidelity computer-based simulation (programming and visuals that get the message across, but does not closely match the real thing) |
| 8 | Multiple-choice "simulation" (items organized into one or more cohesive scenarios using audio/video clips and some computerized simulation; despite their inherent interdependence, items are designed to be as independent as possible) |
| 9 | Multiple-choice situational judgment test (with or without audio/video clips) |
| 10 | Multiple-choice test (using audio/video clips and some computerized simulation) |
| 11 | Multiple-choice test (using simple visuals such as photos, figures, and graphs) |

*Note.* "Multiple-choice" assessments may use other objectively scored response formats such as matching and "drag and drop." Table adapted from D. Rosenthal, C. Sager, and D. Knapp (2003). *A Strategy to Produce Realistic, Cost-Effective Measures of Job Performance.* Alexandria, VA: Job Performance Systems, Inc.

### *Getting it Done*

One of the problems with the current ODARS Program is that the Army relies heavily on Army personnel (military and civilian) to carry out the job analysis work. Although TRADOC guidance specifies involvement of a training developer with experience in job analysis, it is unclear the extent to which most occupational analyses are implemented with such support. Although guidance is provided through Army regulations, there is no expert oversight to help ensure sufficient quality of the resulting work. Moreover, it is not easy to do this type of work well. Like anything else, it is best to involve those who have been specifically prepared for the requirement. We therefore recommend that, although proponents should retain responsibility for and support the collection of occupational analysis information, the work be conducted with the support of professional job analysts. This support could be provided through ARI (there is historical precedence for this) or obtained through contractor support.

## Chapter 4. Test Development

Once the assessment is conceptualized (in terms of form and content), it must be constructed with the input of job experts. In the Army's new assessment program, it will be necessary to have multiple forms of each assessment because it will not be practical to have all Soldiers take the exams at one time on a single day (as is the practice for the Navy). Fortunately, it is possible to construct computerized "banks" of test items that can facilitate development of alternate forms.

### *Challenges*

There are two challenges of particular note for the Army with regard to test development and maintenance. The first regards the requirement for SMEs to participate in the process. It is possible to use (and in fact, we recommend using) contractor support to handle as much test development work as possible. This is likely to result in better products while minimizing involvement of military personnel who are already overburdened. It is likely that professional test development companies will respond to the Army's need by hiring recently separated Army personnel to satisfy this requirement. Models for this are found in the training development fields where professional contract firms work closely within the military community to develop doctrine and materials to meet the requirements of the various proponents.

Another challenge relates to the use of assessment methods that are not typically found in large-scale operational assessment programs. The testing industry has a great deal of experience with running large-scale programs that involve relatively traditional multiple-choice examinations. These methods can easily be adapted to those examinations (like the ones we recommend for the Army) that include variations like drag and drop items and the liberal use of visuals (e.g., photos, figures, and videoclips). There is less precedence, however, for the efficient maintenance of programs involving situational judgment tests, hands-on tests, and high fidelity computerized simulations. Managerial assessment centers are an exception to this, but assessment centers are not often used to frequently reassess the same group of examinees. The point here is that the Army is treading ground over which it had trouble before (at least through the hands-on testing portion of the SQT program) and for which the civilian world and the sister services will have fewer lessons learned to share.

### *The Basic Process*

Figure 4 outlines a generic process for developing and maintaining test materials. To illustrate, we will step through the process as it applies to knowledge-based multiple-choice examinations. Multiple-choice situational judgment tests, in which correct "answers" are based on expert judgment, require additional steps to develop scoring protocols.

In Step 1, test questions (i.e., items) are developed. There are many guidelines associated with the development of high quality items (e.g., framing the question in terms of what is required in a job situation rather than a textbook, avoiding "all of the above" options). Professional test developers facilitate the process by training SMEs to write effective items, then reviewing and editing the items as a quality control measure. This is a skill that requires both training and experience – it is hard for an SME to prepare high quality items the first time around. As mentioned, test development companies will often hire SMEs and thoroughly train them on how to develop good questions.

Development

Step 1. Draft test items/materials
Step 2. Develop scoring protocols
Step 3. Review & revise items/materials
Step 4. Field test items/materials
Step 5. Create test forms

Maintenance

Add item statistics to item/test bank
Review item/test bank for currency

*Figure 4. Major test development and maintenance activities.*

Step 2 is quite straightforward for traditional multiple-choice questions. Item developers are required to identify a source (e.g., technical manual) that justifies the selection of a particular response option as the correct choice. In the item review and revision process (Step 3) other SMEs review the draft items and revise them to make them better. On a professional quality examination, most items have been revised multiple times to improve them before they are administered on an operational examination. A particular area for concern in early reviews of draft test items is the possibility of multiple correct answers.

Ordinarily, several SMEs are involved in the development and review of test items (Steps 1-3). This helps ensure the resulting items reflect the diversity of knowledge, skills, and experience across experts. Each year, the Air Force brings together roughly 600 SMEs (18-20 per Air Force Specialty) at Randolph Air Force Base for a period of up to 30 weeks each to develop test items (see Appendix C). They are provided several days of training and work with a member of a team of civilian test developers to write and review new test items. The Navy generally brings in one (sometimes two) SME for each job (rating) (see Appendix D). This is a 3-year assignment, so the SME has a great deal of time to learn about effective test development (including a 4-day formal training program), but the resulting assessment may unduly reflect that individual's unique experiences and judgment.

Step 3 is typically completed with the collection of "content validity" ratings for each new test item. The content validity ratings are ideally made by an independent group of SMEs who rate each item with regard to its relevance to the job, criticality (to avoid items that cover trivial information), and pertinence to the test blueprint/specifications.

No matter how thorough the SME review and revision process is, unanticipated problems arise when the items are administered to real examinees. In Step 4, items are administered to examinees and the resulting item statistics are used to determine their final status – ready for operational use, need to revise, need to delete from the bank. Ideally, the field test of new items is embedded with administration of the operational assessment. For example, a 100-item exam might routinely include an additional 20 field test items (randomly distributed throughout the test) that are not used as a basis for calculating scores. Neither the Air Force or Navy explicitly field tests items, but they exclude clearly flawed items from the final test scores on each exam. That is, the final score on a 100-item exam may be based on just 95 items because five items may have had unacceptably poor item statistics.

In Step 5, test forms are developed for administration. For multiple-choice exams, this involves drawing items from the item bank that conform to the exam blueprint. There will be items that have been used before (on the immediately preceding form or in previous years) and newly field tested items. By testing all eligible personnel on the same day, the Navy simplifies things because they only require a single test form. Moreover everyone is tested every year, so there is no need to be able to compare scores on a test given one year to scores on a test given a year later[1]. As discussed in Chapter 5, however, this is not a practical solution for today's Army. Therefore, multiple test forms will be developed for each annual assessment period and they will need to be equivalent across assessment periods.

There are at least two strategies for handling multiple test forms that will be considered for the Army's multiple-choice assessments. The first involves creation of 3-4 equivalent forms that will be statistically "equated" to each other and to preceding forms. This method requires that each form of the test contain a subset of items that are common to the other forms. The second strategy makes use of a psychometric method (item response theory) that allows calculation of equivalent test forms with a wide combination of items. It even allows development of equivalent test "forms" that are unique to each examinee. There are limitations to this strategy (e.g., it cannot be used for small MOS exams because it requires large numbers of examinees, and it can be difficult to ensure the blueprint specifications are adequately met), but for some applications it will be very useful.

Figure 4 does not include a step to set a cut score for each assessment that is required for establishing pass/fail cutoffs. Since the Army intends to use test scores as a basis for awarding promotion points, it will not be necessary to set a pass score on the assessments. This is a good thing, because establishing pass scores is a very difficult process that is frequently done incorrectly. It is much better not to have to do it. We recommend awarding promotion points in a manner that requires a simple statistical transformation of test scores (that might range from 0 to 200, for example) to promotion board points (that might range, for example, from 0-150).

### *Other Assessment Methods*

There are significant variations on the development process required to support other less traditional assessment methods. For situational judgment tests, content for new test items is generated through workshops with job incumbents. The process of developing a scoring protocol

[1] Test items vary in difficulty, so it would not be fair to randomly select items from an item bank to create unique a test for each Soldier. That is, some Soldiers would get a harder test than others unless one statistically alters the scores on the different forms to make them comparable.

(Step 2) requires access to SMEs to judge the effectiveness of each response option for each new item. At this point, there is no generally accepted method for generating equivalent forms.

Hands-on tests that are highly proceduralized (i.e., there is only one approved way to perform the task) are relatively straightforward to develop and the Army is quite experienced with this method. For assessment (as opposed to training) purposes, however, more attention must be paid to standardizing administration procedures and conditions and scoring protocols. Such tests would not require multiple forms. Rather, Soldiers might be tested on alternate sets of tasks if a method for ensuring equivalent difficulty across task sets could be devised.

High fidelity computerized simulations are particularly useful for jobs in which key elements of the job can be simulated with a computer (e.g., air traffic controllers, computer programmers, pilots, various types of system operators). To develop sophisticated computer-based simulations requires considerable time with individual SMEs to define the underlying variables in situations to be simulated and ways the computer program should respond to actions by the examinee. Computer simulations are generally expensive to produce, but they are highly standardized, easy to administer, and can quite realistically and safely simulate critical job tasks.

### *Roles and Responsibilities*

In the demonstration assessment program (DCAP), test development is being done primarily through contractor support under ARI oversight. This responsibility is likely to shift to the Army Training Support Center or some other organization and MOS proponents under the guidance of whatever entity is responsible for the overall management of the Army assessment program. As with the job analysis work discussed in the preceding chapter, however, it will probably work best to rely on professional test developers obtained through contractor support to handle the bulk of the requirement.

# Chapter 5. Test Delivery

By the time the Army's competency assessment program is fully implemented, it will involve delivering assessments to tens of thousands of enlisted Soldiers each year. Moreover, some assessments will require different delivery methods (e.g., hands-on versus multiple-choice tests). The advantages of delivering tests on computers via the web are fairly obvious. Managing and maintaining security while shipping paper tests to and from sites all over the world is a monumental, time-consuming task. Tests must be ready to go long before the scheduled administration period and cannot be changed once they have been printed in volume. Loss of even one test booklet compromises the entire process. Therefore, we assume that as much of the test delivery as possible needs to be automated.

## *Challenges*

The test delivery system needs to reach Soldiers in the active and reserve forces. It needs to reach Soldiers who are deployed or the assessment system will not work for purposes of promotion testing.

The biggest challenge to computer-based testing in a large-scale program is having an infrastructure for delivery – that is, secure facilities set up with suitable computers and monitored by test proctors. In any high stakes assessment program, examinees will be motivated to score as well as they can. It would certainly be possible to deliver tests over the Internet to Soldiers just about anywhere, but it would compromise the integrity of the program to do so. The Army needs to know that the right Soldier is taking the assessment and that the Soldier has no unfair advantage over other examinees (e.g., in terms of prior knowledge about what will be on the test or access to resource materials during the test). This is an assessment program, not a training exercise.

## *Available Resources*

Ironically, it is the infrastructure the Army has set up to support training – the Digital Training Facilities (DTFs) – that offers the most promise for handling most of the test delivery needs associated with the new competency assessment program. These facilities are well-equipped, but currently underutilized.

There is a relatively small but nonetheless important precedent for this idea, which is a new assessment for selecting Army recruiters that is being delivered via a subset of the DTFs. The biggest challenge faced in implementation of this assessment through the DTFs was arranging for suitably trained personnel to proctor the exam at the DTFs. This will be a much larger requirement for the competency assessment program. In the short-term demonstration period, however, proctoring will be handled primarily by research personnel provided by ARI. A detailed discussion of the DTFs and the requirements associated with web-based delivery of Army competency assessments via the DTFs is provided in Appendix E.

### *What the DTFs Will Not Be Able to Handle*

The DTF solution, however, may not support administration of high fidelity computer-based simulations (which could conceivably require special response pedestals or other non-standard computer capabilities) and will not support hands-on tests. This latter point is of concern because hands-on tests may be a highly desirable component of testing for a large number of MOS. In Phase I, we have not focused on this problem, but will need to do so in Phase II when prototype MOS-specific tests will be developed. These assessments will almost certainly include hands-on measures or other test methods that cannot be delivered via the DTFs.

It is possible that practical considerations make it infeasible for the Army to include any assessment methods that cannot be delivered via DTFs. In the PerformM21 project, however, we will explore ways to make this a feasible option for at least some MOS (e.g., those with relatively low numbers of Soldiers). For example, it may be possible to train NCOs to administer hands-on tests in their units. We know that this can be done – the Army already administers Common Task Tests in the field every year. What will be required for this to work for purposes of more high stakes individual assessment, however, is greater standardization of the assessment process, better record keeping, and a commitment to administer assessments with an assessment (as opposed to training) focus. This is hard for NCOs because they are trained to be trainers rather than assessors.

# Chapter 6. Interfacing with Soldiers

If only because of their number, it is no small requirement to communicate with the Soldiers who will be taking assessments each year. Telling them about the assessments, scheduling them for testing, providing score reports, and answering questions is a major undertaking.

Following are some of the tools that the Army should develop and use to help address this requirement:

- Soldier test preparation guides
- Information about the program posted on the web
    - Test dates and locations
    - How the assessments are developed
    - How the assessments are scored and what type of feedback is provided to examinees
    - Quality control measures to ensure tests are fair and accurately scored
    - Frequently asked questions
- Soldier feedback reports

As discussed elsewhere in this report, misinformation and misunderstandings about an assessment program can be very damaging. It is also true that it will not be possible to completely prevent this from happening and that it is human nature to disparage tests that one has to take. Nonetheless, there is much the Army can and should do to promote open and honest communications with Soldiers and NCOs about the assessment program.

*Soldier Test Preparation Guides*

In civilian certification programs, a test preparation or study guide typically includes information about the following areas:

- What is on the test (e.g., a test outline or blueprint)
- How many items are on the test and what they look like (often sample items are included)
- A bibliography of sources that would be useful for studying in preparation for the test
- What the test experience will be like (e.g., test locations, time limits)
- Hints for taking the test (e.g., there is no penalty for guessing so do not skip any questions)
- Warnings about cheating (e.g., that scores may be invalidated)
- How the test will be scored
- What type of feedback will be provided and when it will be provided

The test preparation guide may be completely automated or it may refer to a website for further information or a sample test. Particularly when an assessment is provided in a novel format (as at least used to be the case with computer-based testing), a sample test that familiarizes the examinees with the test administration format prior to test day is highly desirable. It is important that test preparation guides be accurate and provided to examinees in sufficient time to allow them to prepare for the examination.

*General Information*

Although the test preparation guide should be fairly complete, it should be supplemented with a website and other strategies for communicating information and addressing frequently asked questions about the assessment program. The other services use websites and traveling "road shows" to inform enlisted personnel about their assessment programs. Varying the way in which information is presented (in-person, paper-based materials, web-based materials, text-based presentations, oral presentations, graphic presentations) helps to ensure everyone really understands the message. It also helps to keep the message simple. It is not uncommon to make certain program design decisions (e.g., how scores will be calculated) in part on how simple the explanation of the process can be made for examinees.

*Soldier Feedback Reports*

Typically, examinees would like as much feedback as possible on an assessment, including which individual items they answered incorrectly. Providing such detail is not feasible in most operational testing programs because the test items get re-used on future test forms. It generally is possible, however, to provide an indication of how well the examinee performed in various areas covered by the assessment. The key is making sure scores provided to examinees are statistically reliable, so it would be fine to report a subscore based on 20 test items but a subscore based on four items would not be informative.

Examinees intuitively understand "number correct" scores (e.g., "I got 65 out of 100 items right"). Unfortunately, once we have multiple forms of an exam that are likely to vary somewhat in difficulty level, it is necessary to statistically transform scores from the different forms to make them comparable to each other. The good news is that this is possible and it is commonly done (e.g., with college entrance exams). The process, however, can be difficult to communicate to examinees.

Examinees also want to know how they performed relative to others. Therefore, we envision a Soldier feedback report that includes the following elements:

- Standardized scores – total and for each major component of the test)
- Norm-referenced scores (e.g., percentile scores) – total and for each major component of the test
- Bar graphs to depict the profile of strengths and weaknesses as indicated by the subtest scores and to show performance relative to the Soldiers peers
- Text descriptions of how to interpret the information provided and where to go with questions

In Phase II of the PerformM21 effort, we will mock up feedback reports for multiple-choice assessments and consider how to adapt the model to other types of assessments that will be administered (e.g., hands-on tests and computer simulations). It will also be necessary to determine exactly how scores will be delivered to Soldiers and their commanders.

## Chapter 7. Integrating the Assessment Program into Army Systems

An Army competency assessment program will be a major undertaking for the Army requiring change, invention, and adaptability. It is crucial that, from the start, the assessment program not be viewed as a stand-alone or separate ancillary program. It must be fully integrated with other existing programs.

Although it is impossible to completely define the operational climate in which a future assessment program must operate, it is possible to outline a vision for future requirements. In defining the areas of impact, we relied on three primary resources: First was the experience and approaches used by the other services (Navy, Air Force) in their testing programs. Second, was the Army's considerable history and experience with testing, primarily in the form of its SQT program. Finally, is the ATPAT. This group of senior NCOs, representing many different commands, organizations, and experiences, provides a contemporary and broad view of the many operational considerations and implications attendant to adoption of a testing program. We have identified four broad areas of integration requirements:

- The Command and Cultural Climate
- Organizational Structure
- Personnel Policies
- Education and Training

### *The Command and Cultural Climate*

The Army has been without any type of formalized Army wide testing system since the early 1990s; probably 90% of the current force has had no first hand experience with an Army testing system. Start-up and implementation of a test system will have a profound effect on many aspects of the Army command and culture. Foremost, the testing system must receive recognition and support from the highest command levels if it is to become viable, supported, and accepted by the junior enlisted and those within the NCO ranks who it will impact the most. It is significant that the impetus for the revival of the Army testing program came from the ranks of the NCOs – the 35,000 Soldiers who participated in the ATLDP were quite explicit in their description of what is required. Endorsement of their findings and a continuing commitment to implementation at the highest levels will be the first of several required senior Army leadership interventions needed to facilitate the program.

Key to acceptance by the Army is understanding the system and key to this understanding is knowledge. Soldiers must perceive testing as fair and equitable and all aspects of testing must be open and transparent. Long before the system is implemented there needs to be a program that explains and publicizes the system. Lack of coherent and consistent information leads to confusion and the substitution of rumor and legend in place of fact. The Army needs to take advantage of existing Public Affairs, Strategic Communications, and other troop information channels to publicize all aspects of the program. Moreover, the Army needs to establish a centralized testing website where Soldiers can go not only to find information but to also ask questions and seek clarification and receive personalized, real-time feedback. Finally the Army needs to aggressively push the program though spokespersons that travel to the field and explain the program, face to face, with Soldiers. Erroneous beliefs were a major factor in the demise of the SQT but the fact that they were erroneous is immaterial – the important factor is that they were believed. A modern program, based on solid, stable policies, and communicated openly and through various channels and taking maximum advantage of email, websites, and other technology, is an essential piece of the Army test system.

*Organizational Structure*

An Army assessment program must reside in an organization whose sole function is assessment. Quite likely, this will require a new organization, chartered, staffed, and resourced to carry out the Army testing mandate. The new organization would have responsibility for Army testing doctrine and policy and implementation of the testing program throughout the Army as well as overall test administration and utilization policy. It would be directly responsible for the development and maintenance of the Army-wide competency assessment test. The group that assumes the follow-on duties to the ATPAT, (the Test Council of SGMs, see Figure 1 in Chapter 1) should operate under the control of this organizational entity. This Army organization should be at the Directorate level, headed by a Colonel (O6) and sufficiently staffed with both test psychologists and military representation.

This organization would likely be part of a major command or directly under the DA staff. Its organizational location may depend on the primary focus of the Army testing system – a focus on use as a promotion tool (personnel) or its training focus. Ultimately, this is an Army policy issue; however, there is considerable evidence that a promotion tool focus is more compatible with Army goals than a test with primarily training goals. A critical lesson learned from the SQT was the extraneous and distracting emphasis that commanders experienced when training and readiness indicators became a major consideration of the SQT. As a promotion tool, an Army assessment program would still have major attendant training and readiness effects but, it is hoped, without the adverse consequences experienced with the SQT.

A further organizational implication and consideration focuses on the alignment of the assessment function with the occupational analysis function. Operationally, the two requirements are closely intertwined, although occupational and job analysis also plays a role in other training and job definition needs as well. In the other services, the Air Force combines the two functions within the same organization, even physically co-locating them. The Navy employs a more remote, although symbiotic, relationship. The introduction of a measurement function should cause the Army to reexamine its occupational and job analysis function as well including evaluation of the organizational structure for both. The office responsible for the assessment function will become a primary user of the products of the job analysis; if not organizationally joined, the two functions must have a clearly established operational relationship.

*Personnel Policy*

An assessment system used as a promotion tool will obviously require reexamination of all existing promotion policies. The Army uses basically two different NCO promotion programs: The semi-centralized promotion system to ranks of Sergeant (E5) and Staff Sergeant (E6) and the centralized promotion system which promotes to Sergeant First Class (E7), Master

Sergeant (E8), and Sergeant Major (E9). The semi-centralized system is a points based program which (currently) uses a local board points award system, administrative points, and the commander's points recommendation to determine promotion standing within the individual Soldier's MOS. The centralized system is less formulated and involves selection by centralized boards based on record reviews and following specific instructions prepared for each selection board.

The semi-centralized system is the easiest to revise and is the obvious start point for the introduction of a test component as a part of the promotion system. It is also a system that is primed for review and revision, which, except for some superficial adjustments of points emphasis, is basically the same system that existed in the 1970s. The centralized promotion system is a more complex issue. Because it is less structured, the introduction of promotion test criteria needs more review and consideration as to how it should fit within either the existing or a revised centralized promotion system. One fairly simple adoption strategy would be to just provide test results as another element for the selection board to consider.

A consideration for the semi-centralized promotion policy implementation is the probability that individual MOS may have different job testing policies. As long as promotion decisions continue to be made by MOS, this should not be a major issue. Even the fact that some MOS may not have job-specific tests at all does not have fairness or equity effects if handled properly. However, it is critical that the policies be clear-cut and in place and publicized before the testing program is implemented. Ancillary personnel issues that need to be addressed and programmed include test recording, reporting, and record keeping issues and the interface of test results and individual personnel records.

The translation of test results to promotion decisions involves issues of scoring, management, weighting, standardization, and equating. There is a great deal of flexibility possible in the application of test results as well as many considerations and cautions in how testing results are applied. It is crucial that personnel policy decisions be based on sound incorporation of testing science and experience and that there be a mutual understanding of both the requirements of the promotion program and the products and application of the test program, melded into personnel policies.

A final area with personnel policy implications is that of test security. Both the Navy and Air Force have clear-cut personnel policies and command emphasis on test compromise, test control, and cheating. Both services investigate violations aggressively and have successfully prosecuted individuals under the Uniform Code of Military Justice (UCMJ) for willful or deliberate breaches. The Army needs to identify, implement, and publicize a similar approach.

### *Education and Training*

Even a testing program with a primary focus on promotion will have a significant impact on training, training materials, and training delivery systems. Self-development programs must be in place that match and support the test effort, both for the Army-wide generalized Soldier competency areas and for any MOS-specific evaluation that eventually becomes part of the system. Test developers, trainers, and Soldiers must all work from the same resources and references and accessibility of materials is a major consideration for all. While the ATPAT expressed the principle that the test system not, in itself, generate the production of new training materials, this area must be constantly reevaluated to ensure that needs and products are matched.

One area of critical interface is the Non-Commissioned Officer Education System (NCOES) and in particular the Primary Leadership Development Course (PLDC) and the Basic Non-Commissioned Officer Course (BNCOC). PLDC is a standardized, 4-week, common core resident course taught at various NCO Academies (NCOA), both in CONUS and OCONUS. PLDC is a prerequisite for promotion to Sergeant (E5). BNCOC is MOS-specific, varies in length by the MOS, and is normally a resident course at the MOS proponent location. All BNCOCs have a common-core component. BNCOC is prerequisite for promotion to Staff Sergeant (E6).

A test system and NCOES share too many common interests and threads not to be closely integrated. Both have the shared goal of preparing and qualifying the Soldier for promotion. As the testing program is developed and matured, the goal should be to maximize integration of the two programs, both doctrinally and administratively. This would be beneficial to both programs and could eventually lead to cost savings in the administration of the testing program.

# Chapter 8. Program Evaluation

As the Army's new assessment program is implemented, problems will inevitably surface and require modifications to program procedures and policies. It would be to the Army's advantage to proactively look for these problems and areas that might be improved. Through program evaluation, problems can be identified and addressed before damage occurs. We recommend that evaluation strategies be incorporated into all phases of program implementation and become a permanent part of the program. Not only would this be in the Army's best interests, but it would also address provisions of the Government Performance and Results Act (GPRA) as recommended in a recent U.S. General Accounting Office review of TRADOC activities (GAO, 2003).

We also recommend managing the expectations of program stakeholders (e.g., Soldiers, proponents, test developers) so they appreciate the willingness of and need for the Army to change the program to be successful over time. It will not be a static entity that, once in place, will not change.

## *Evaluation Methods*

Program evaluation can take many forms. Researchers typically speak of formative evaluation and summative evaluation strategies. Formative evaluation generally involves obtaining stakeholder (e.g., examinees, test developers, test delivery personnel) reactions to a program. It is usually fairly easy to incorporate formative evaluation processes as an ongoing feature of a program (e.g., to routinely solicit feedback on test items from examinees whenever an exam is given). Summative evaluation is more demanding, in that the goal is to provide empirical evidence that the program is meeting its intended goals. For example, it might require a research project in which assessment test scores are correlated with subsequent job performance at the next pay grade.

Concerns that should be explored in program evaluation efforts include the following:

- To what extent are the assessments reliable and content valid?
- Are unnecessary costs (in terms of money and time) being incurred?
- Are support contractors performing adequately?
- Are there inefficiencies in how the program operates across MOS and supporting offices?
- Are the Soldiers who perform best on the assessments the same Soldiers who perform the best on the job?
- Is the assessment program promoting more effective training and self-development?

## *Managing Expectations*

Closely related to the idea of program evaluation is the issue of stakeholder expectations. This is particularly true in testing programs because those who have already been tested tend to resist changes to the program that might be viewed as making it easier for future examinees. Thus, they expect that once the assessment program is in place, it will become the standard and will not change. Those responsible for running the assessment program may also resist change because it requires that they change how they do business. This type of reaction ties the hands of those interested in improving the program as lessons are learned and new technology offers new solutions.

# Chapter 9: Conclusions and Next Steps

## *PerformM21 Phase I*

At the end of Phase I of the PerformM21 research program, we have a general idea of what a new Army competency assessment program might look like and we have a prototype core assessment targeted to Specialists/Corporals (E4s) eligible for promotion to Sergeant (E5) ready for pilot testing. We have learned nothing so far to indicate a new assessment program is not an idea worth continued serious consideration. It is also clear, however, that the Department of the Army needs to understand the scope of commitment required to support such a program before making the final decision to do so. This includes commitment of substantial resources, communicating with stakeholders to address their concerns and obtain their cooperation, and commitment to quality so that Soldiers are subjected to tests that are both valid and fair. That said, if the Army is successful with a new assessment program, it will have successfully responded to key recommendations from the ATLDP panels and the outgoing SMA. Moreover, the positive effects of the program will be evident in improved force readiness. This is, after all, the ultimate goal of the program and what most makes it worth pursuing.

## *A Look Ahead to Phase II*

In Phase II of the PerformM21 project, researchers will administer the prototype core assessment to a sample of Specialists/Corporals. We will use their test data to evaluate and improve the test questions. We will also ask participating Soldiers for feedback on the entire assessment process (e.g., signing up to take the test, usefulness of the test preparation guide) and use this to improve elements of the recommended operational program.

We will take the prototyping part of the research a step further by designing and developing job-specific assessments for several MOS. This will allow us to tryout different job analysis strategies and different assessment methods (e.g., computer simulations).

While all this is taking place, we will continue to work with the ATPAT and other resources to develop more detailed recommendations regarding the design of an operational competency assessment program. These recommendations will cover aspects of test design, development, delivery, and maintenance, as well as considerations related to policy development and management of the program. We will endeavor to provide enough information about program requirements to help the Department of the Army determine the financial costs of adopting the program (in all or in part). The ultimate goal of the project is to help the Army make fully informed decisions about a major new initiative.

# References

Department of the Army. (2002). *The Army training and leader development panel report (NCO).* Final Report. Fort Leavenworth, KS, U.S. Army Combined Arms Center and Fort Leavenworth.

Campbell, R.C., Keenan, P.A., Moriarty, K.O., Knapp, D.J., & Heffner, T.S. (2004). *The Army PerformM21 Demonstration Competency Assessment Program development report* (IR03-105). Alexandria, VA: Human Resources Research Organization.

U.S. General Accounting Office. (2003). *Defense Management: Army Needs to Address Resource and Mission Requirements Affecting its Training and Doctrine Command* (GAO-03-214). Washington DC: Author.

Rosenthal, D., Sager, C.E., & Knapp, D.J. (2003). *A strategy to produce realistic, cost-effective measures of job performance.* Alexandria, VA: Job Performance Systems, Inc.

# APPENDIX A
# PerformM21 Needs Analysis Organizing Structure

*Note the test program must be applicable for the reserve components (Army Reserve and National Guard). Also test security is not treated as a separate category since it will factor into many of the design and policy decisions noted here.*

Purpose/goals of the testing program

- How will it be used (currently, primary objective is to serve as a promotion test)
- What effects is it supposed to have (e.g., improve force readiness, encourage self-development)

Test content

- Underlying philosophy (e.g., driven largely by job analysis with SME modifications; representative sampling of relatively important content versus targeting areas likely to discriminate among candidates)
- Procedures for generating and updating job analysis data
- Procedures for generating and updating test blueprints
- Procedures/strategies for minimizing the need for having alternative test "tracks" (e.g., for different types of personal weapons)
- Relationship among test content (i.e., target areas, individual test questions) across skill levels, Army-wide versus MOS-specific exams, and test administrations. (Note that content equivalence must be maintained across candidates in same "pool" but not necessarily for candidates not being compared to each other.)

Test design

- Psychometric characteristics (e.g., classical test construction, Item Response Theory (IRT); desired difficulty)
- Test format and length
- Handling test tracking requirements, including ensuring equivalent scores across tracked tests
- Delivery mode (currently, Army wants web-based)
- For given administration cycle, how many test "forms" required?
- Procedures for updating and developing new forms (including schedule for doing so)
- Reflection of performance standards (e.g., cut scores), if applicable. Currently, don't anticipate need for pass/fail cut point, but consider strategies for conveying performance level to candidates (e.g., by norm group comparisons or identifying broad performance categories tied to test content/difficulty)

Test development

- Reference sources (need to be available and up-to-date; how will this be accomplished?)
- Item writers, reviewers (e.g., Advanced Individual Training (AIT) instructors or field NCOs? how many are needed for how long? use contractors to supplement?)
- Role of technology to minimize resource requirements

- Pilot testing (e.g., always have "overlength" forms so poor items can be dropped without sacrificing blueprint content specifications)
- Item banking procedures and supporting software

Test administration

- Test administrators/proctors
- Locations/facilities
- Frequency
- Identifying and scheduling eligible candidates

Interfacing with candidates

- Test preparation (other than that covered under self-assessment) – informing candidates in a timely fashion about test content and study references
- Amount and type of feedback on test performance
- Development and provision of test preparation guides

Associated policies

- Eligibility requirements
- Retesting; make-up testing
- Candidate rights (e.g., to challenge individual test questions)
- Test preparation (e.g., encourage or forbid group study)
- Rules for allocating Promotion Board Worksheet points

Links to various Army systems

- Possible adjustment to promotion system rules and timing (currently monthly promotions) to facilitate smooth integration with test program
- Establish clear links to training systems and philosophy (e.g., focus on procedural knowledge versus underlying principles, consistency in study references and training content)
- Smooth integration of testing history/results into personnel records and any other required reporting or archiving needs

Self-assessment

- Explore strategies for supporting link to training needs
- Explore strategies for integrating with self-development (e.g., distance learning) requirements
- Identify link between self-assessment tools and competency test content and procedures
- Identify potential methods (e.g., sample test with feedback on each question; self, peer, and/or supervisor ratings; video-based problems with associated discussion; practice hands-on tests with buddy or unit support/scoring). Note that last example here suggests we might be able to extend "self-assessment" to include anything other than "for real" testing.

# APPENDIX B
# Skill Qualification Tests: Brief History,[1] Lessons Learned, and Recommendations

As efforts to design and implement a performance assessment system are underway, it is instructive to examine prior efforts to identify both their strengths and their weaknesses. The goal is to develop a system that accurately describes an individual Soldier's level of technical job competence, is feasible for standardized administration, and allows a fast turnaround of feedback to Soldiers, units, and the Army to support training and leader development programs and improve Soldier and unit readiness. These were also the goals for the Army's Skill Qualification Test program in the 1970s and early 1980s, yet the program went through almost constant changes and eventually collapsed under the untenable weight of administrative requirements.

Thus, re-instituting a new performance assessment system must address both the factors that led to creation of the SQT program and those that forced its abandonment. How will the new system build on the legacy of the SQT and take advantage of technological advances that can reduce the developmental and administrative burdens encountered with SQT?

## An Historical Review

### *The Initial Years: 1977-1980*

From the late 1950s through 1976, the assessment system was based on the Enlisted MOS Evaluation Score. The principal components were a written examination and the supervisor's rating of the Soldier's job performance (some MOS also had a performance component, such as a typing test or playing a musical instrument). The written test items tended to be knowledge-oriented, addressing equipment nomenclature, regulations, publications, and knowledge of rules. The concept of task-based testing was not yet common, and few items were procedural. For the most part, technical difficulties in reproducing drawings and photographs precluded all but textual items.

It should be noted that the MOS tests were used almost exclusively for *personnel-related activities.* Trainers and training developers were not involved, nor were unit commanders. The purpose of the test was to inform promotion and selection activities.

In the early 1970s, General William Westmoreland, the Army Chief of Staff, began to place an increased emphasis on training and to decentralize planning and conduct of training to the platoon and squad levels. The U.S. Army Training and Doctrine Command (TRADOC) was established in 1973, with General William E. Depuy as commander and Major General (MG) Paul F. Gorman as his deputy for training. Both were advocates of performance-oriented training, and recognized the need for an assessment system that would support both training needs identification and personnel actions. In particular, MG Gorman brought a new concept of a

[1] Prepared by Charlotte H. Campbell. Adapted from Campbell, R. C. (1994). *The Army Skill Qualification Test (SQT) Program: A Synopsis* (HumRRO Interim Report IR-PD-94-05). Alexandria, VA: Human Resources Research Organization.

systemized way to set training objectives through the careful determination of tasks to be trained, conditions under which certain training would be required, and the setting of standards. TRADOC proposed that the Army needed a program that would require Soldiers to perform to established standards. In addition, the program should be progressive and sequential, so that each level of objectives built on the next lower level.

The proposal to implement performance testing for enlisted MOS was approved by the Department of the Army in 1974, and for the next 3 years there was intense activity to design the system for analysis, development, implementation, and evaluation. It was intended from the start that the SQT be a training management and evaluation program. While it would be used for purposes of personnel management, the testing fell under the authority of the training management function, and was generally implemented within units (usually at the battalion level). The SQT was not a stand-alone system, but rather was intended to dovetail with the new enlisted job and career structure program, the Enlisted Personnel Management System.

The SQTs assessed occupational proficiency in each job (MOS and skill level), based on analysis of critical skills required for proficiency in that job. The concept of the *task* was central to the structure of the SQT: the test's scoreable units (SUs) were oriented on tasks, and the final score for a Soldier was expressed, generally, in terms of SUs on which the Soldier demonstrated proficiency. The testing method was a combination of a hands-on performance evaluation, written test, and supervisor's performance certification of specific tasks. Every test contained SUs targeting the tasks for that skill level, as well as for the next higher skill level, so that the tests could be used both to *verify* proficiency at that skill level and to *qualify* the Soldier for promotion to the next skill level.

The hands-on component (HOC) of an SQT was the high-visibility element. In general, the HOC was regarded as exceptionally valid, in that it required actual task performance under high-fidelity simulated field conditions, with scoring by a trained observer on both process and product measures. The scoring record was limited by the technology: manually-recorded scores were transferred to machine-scored sheets (with the attendant opportunities for error), and the machine-scored sheets could only handle 20 or fewer performance measures per SU. The testing was resource-intensive, requiring considerable preparation, use of actual equipment, field time, and a large cadre of administrators. There were also provisions for units without adequate resources to administer only part of the HOC for an MOS, or to alibi the entire HOC.

The written component (WC) was multiple-choice, and like the HOC was organized around tasks. Each task-SU could have between 3 and 10 items, items could have up to 10 alternatives, and there could be more than one right answer. Unlike SUs in the HOC, the WC required that proponents establish passing standards (less than 100% correct, in most cases) for each SU; the intent was to recognize that written measures were inherently less valid than full performance measures. The guidelines were complex, considering the number of items, number of alternatives per item, and number of correct alternatives. Like the HOC, the WC was also limited by the requirement to use machine-scorable record sheets.

The performance certification component (PCC) was never a winner. It was designed to test tasks that would normally be assigned to the HOC, but could not be tested there because equipment needs, conditions, or time requirements made it unfeasible. It was intended that the PCC be administered by a supervisor during normal working conditions or other certifying events (e.g., weapons qualification, physical fitness testing) during the year.

All three of the components had provisions for *tracking* – allowing Soldiers to be tested on some different tasks or portions of tasks if they routinely worked on different equipment types or had specific duty positions.

To be eligible to take an SQT, Soldiers had to be in the Army for at least a year and have held the MOS for 90 days. However, the guidance on eligibility was characterized by numerous waiver provisions, and because the testing was generally a unit activity, many Soldiers who were not eligible took the SQT, while others who were eligible did not.

The plan was that MOS-proponent agencies (for the most part, the TRADOC schools) would develop the SQTs, and also try them out and validate them. The tests would then be shipped to the Individual Training and Evaluation Directorate (ITED) at TRADOC for printing and distribution to some 60 Test Control Officers located world-wide. Test Control Officers then coordinated with the units for schedule and support. Unit personnel administered the SQTs and returned the scoresheets to the Test Control Officers, who sent them back to ITED. ITED was responsible for scoring, verifying the scores, and reporting results to the Soldiers and to the units. ITED also had responsibility for analysis and research on the aggregate test results.

The SQT became operational for a selected set of MOS in January 1977, at which time the old MOS tests would no longer be administered Army-wide. All MOS (with some exceptions) were to have their SQTs in place and operational by 1980. In fact, while all proponent agencies had some SQT development work underway and many SQTs were implemented by 1980, overall the schedule for implementation slipped significantly and the program was in trouble.

There were three major sources of the difficulties, and they pervaded the entire system. One source concerned the proponent agencies. They were to be responsible for analysis, development, tryout and validation, production of camera-ready versions of tests, administration manuals, and SQT notices to Soldiers. Because of the printing burden on ITED, there was a concomitant burden on the proponents for standardized camera-ready material within tight time schedules. Despite increased staffing and an intensive series of SQT development workshops for developers, proponents were unable to meet the demand for high quality tests for all MOS. The timeline was revised to have all SQTs operational by December 1981; this goal was also not met. A second source of the difficulty had to do with the workload at ITED. Production and distribution requirements were massive, and delivery to the field was not always timely. The quality oversight to the proponents was stretched thin, and some flawed products were fielded, resulting in later invalidation of some SUs and in some cases of whole SQTs.

The third major obstacle to timely implementation occurred in the field. Units were responsible for not only getting the Soldiers to the appropriate tests, but also for equipment and scorers and other administrative support. SQT preparation and administrative activities could virtually swamp all other training and operational action for up to 15 days prior to the testing itself. As the pressure to "do well" within units increased, the time devoted to individual Soldier preparation increased. Perceptions of unfairness were rampant, not only because of the unevenness in administrative support but also because some MOS got their SQTs late or not at all.

While there were problems throughout the system, it should not be inferred that by 1980 the SQT system was in total chaos. Many tests of very high quality had been produced and large numbers of Soldiers were being assessed with great validity. Most importantly, deficiencies (and proficiencies) in individual performance and training were being identified. The concept of "performance-based" training and testing had taken hold throughout the Army, and the SQT was one of the key drivers in that shift.

*Retrenchment: 1980-1983*

The problems perceived by leaders and by the Soldiers themselves had to be addressed, so revisions to the organization, development, and implementation of the SQT were instituted. First, in 1980, the requirement to include some proportion of tasks from the next higher skill level was dropped, along with the concepts of *verify* and *qualify.* A minimum passing score of 60% was set. With several interim solutions, the actual percentage was eventually used in computing promotion points (where number of points equaled double the percentage level), providing the Soldier achieved the minimum score. The SQT could thus contribute between 120 and 200 points to the total promotion points, out of a possible 1,000 points total.

Several structural changes were made to the SQT design. The HOC was essentially unchanged, although new guidance directed (somewhat vaguely) that a typical unit should be able to test all its people in 8 hours. Proponents were tasked with preparing an Alternate HOC – a written test that could replace the HOC if the HOC could not be administered. The Alternate HOC would cover the same tasks as the HOC, and there would be little if any overlap with the WC.

The WC was dropped and replaced by the Skill Component (SC) – also a written, multiple-choice test. However, the test did not have scorable units linked to tasks, which had had the effect of equalizing the weight given to tasks despite the number of items per task. There was to be less reading and more visual presentation in both the questions and the responses. The immediate effect was that all items became four-alternative, one correct answer items. In fact, difficulties in scoring the more innovative 10-alternative items with multiple correct answers had already been noted and most developers had moved to the more standard format. Additional guidance reduced the time allotted for the SC to one hour for combat MOS and two hours for technical and support MOS, and also reduced the weight given to the SC in the overall SQT score.

The PCC was also changed. It was renamed as the Job Site Component (JSC) and became much less test-like. The Soldier's immediate supervisor was provided a list of tasks for that Soldier's MOS and skill level, and could evaluate the tasks any time the Soldier performed them, whether or not the Soldier knew he/she was being evaluated. There was also a provision that if the JSC task could not be observed and assessed, the Soldier could get credit for Army Correspondence Courses or Training Extension Courses on the task.

These changes bought the program some time, but did not erase the problems. If anything, as SQTs were fielded for more MOS, the problems began to intensify. Proponents found that even with experience and with several versions of the tests already in the vaults, manpower demands were extensive. In some cases, proponent in-house test expertise was still lacking, and test quality suffered. Resources at ITED were strained to the breaking point.

The greatest impact was felt in the units, however, and the strongest reactions came from unit commanders. Even with the reduction in the scope of the individual testing, the overall effect was that there was more testing going on, and commanders could easily see the greater demands on resources generated, in particular, by the HOC requirements.

Finally, in March 1982, a General Accounting Office (GAO) report was issued that was highly critical of the SQT system overall. Although that report is sometimes cited as the SQT death knell, and its Congressional focus definitely insured its notice, the examination of an outside agency was probably not necessary. Most of the criticism of the SQT was being generated from within the Army, and the Army leadership was willing to listen and make changes. More importantly, there had been numerous personnel changes at leadership levels, along with the usual policy shifts.

The GAO report identified several shortfalls, including:

- The test results did not accurately indicate a Soldier's ability to perform critical job tasks because only a selected number of tasks were tested.
- SQTs were being used as once-a-year events, instead of being the culmination of a year-round training program.
- Promotion decisions based on SQT results created inequities among Soldiers.
- Test results were not used routinely to measure either individual proficiency or unit training needs.
- The SQT program hampered individual professional development because training was being provided primarily for the limited number of tasks being tested.

The report also addressed the cost effectiveness of the SQT program. It noted that the program had become a "paper nightmare," requiring thousands of people each year to develop, print, distribute, and score the hundreds of tests at an annual cost of over $25 million.

Noting the valid need for a system that measured Soldier proficiency and identified training needs, the GAO report recommended that the two functions be separated. The program for assessing individual training needs should be tied directly to Soldier's Manual tasks and should be used as a training diagnostic tool for individuals and units. Tests of individual proficiency for use in promotion decisions should apply only to those Soldiers eligible for or already within the NCO ranks, and should comprise both written MOS-specific testing and hands-on testing of common tasks.

The Army, in commenting on the report, not only concurred with many of the GAO conclusions, but also stated that growing administrative workloads had led the Army to some of the same conclusions.

*Post-SQT: 1983 and Beyond*

Starting in 1983, the Army officially dropped the full, three-component SQT that had been in place for nearly 6 years (although it was certainly not a constant over that time). In its place was formed the Individual Training Evaluation Program (ITEP). This program retained both the training emphasis and the use as a personnel management tool, although the reporting of results was widely decentralized.

Under the ITEP, the only hands-on testing was in the Common Task Test, which addressed tasks that were applicable across MOS throughout the Army. The JSC was largely abandoned, although there was a Commander's Evaluation that provided for someone in the Soldier's chain of supervision to evaluate the Soldier's ability to perform "mission related tasks" of the supervisor's choosing. There was no reporting requirement, and the results did not in any official way impact the Soldier's promotion chances. The written test alone retained the name of Skill Qualification Test. It remained MOS-specific and task-oriented, and was the only measure of individual proficiency that was reported outside the Soldier's unit.

The changes to the Army's individual evaluation system did not end with the adoption of the ITEP in 1983. The Army dropped the term "SQT" in 1991 and adopted the name Soldier Development Test (SDT) to reflect a philosophic shift: It was felt that NCOs should take more responsibility for their own MOS and their own leadership development. This program too was controversial, and in 1994, TRADOC Commander General William W. Hartzog recommended its elimination, because the Army's Command Sergeants Major felt the test was redundant to NCOES, was not as objective, and did not provide battle focus.

## Lessons Learned

From 1977, when the first SQTs were fielded, until 1991, when the name "SQT" was finally dropped, there were almost constant changes in design, implementation, and use of the tests. These changes, though, were in the nature of evolutionary changes. In contrast, the changes to the Army test system that were adopted in 1977 and remained essentially in place were so bold as to be judged truly revolutionary. Indeed, the core tenets that led to the formation of the SQT program – the notion of individual criterion-referenced job proficiency testing for both training diagnostics and personnel management – are still widely held throughout the military services and are also found in industry and education.

In many respects, the SQT was a victim of the structure and constraints in which it had to operate. Consider the conditions and expectations:

- The program should provide assessment for all enlisted Soldiers – about 396,000 in grades E4 through E6, the primary testing population.

- The program should have job-specific tests for every MOS and skill level – in 1977, there were 180 MOS and most had two or more skill levels – and a significant proportion of the test should be new every year.

- The program should be useful as a training tool – thus, scores should be available to both the Soldier and the unit very soon after testing.

- The program should be useful as a personnel management tool – thus, scores should be maintained centrally to facilitate planning and decision-making.

- Test administration must be fair and equitable for all Soldiers, including those in non-traditional duty positions and those who would be unable to take a hands-on test due to their locations.

Given such expectations, it is surprising that TRADOC, proponent agencies, and units were able to accomplish as much as they did in just a few years. In considering lessons learned, it is instructive and important to consider both what worked and what did not work.

*What Worked*

A continuing strength of the program was the focus on tasks and task performance measurement. Criterion-referenced training and testing were just coming into vogue in the early 1970s, and the SQT program was firmly grounded in the notion that proficiency on critical aspects of the Soldier's job should be measured. The SQT program was only a part of the trend toward providing well-defined job descriptions, comprehensive job and task analyses, and objectives-based training.

Early advocates and planners of the SQT program were not entirely naïve with respect to the difficulties inherent in writing and validating this new type of test. Part of the effort to ensure quality involved the development of a comprehensive *Guidelines for Development of Skill Qualification Tests*. This guide contained comprehensive instructions covering preparation of an overall SQT plan, development of all three test components, conduct of tryouts and validation options, and preparation of final camera-ready materials, along with examples, checklists, and suggestions for problem-solving. Even further, TRADOC sponsored the creation and delivery of a series of week-long training workshops in which nearly 200 personnel were trained as trainers who could then teach proponent-agency developers how to prepare SQTs. Now, 25 years after the workshops were conducted, some of those trained trainers are still at TRADOC schools and still incorporate criterion-referenced testing principles in their work.

Even in the written tests, the emphasis was always on performance. Written SUs were often conceptualized as being "performance-based" when they could not require authentic performance of a task or task step. Developers were encouraged to use novel approaches, which included, at a minimum, the use of pictures in both item stems and alternatives.

*What Didn't Work*

The intent that SQTs be both training tools and personnel management tools made sense in the abstract, but in reality the dual-purpose led to a continuing tug-of-war. As a training tool, it made sense that units have charge of implementation and have timely reports of results. But as a management tool, the scores needed to be added to a universal personnel database in order for promotion decisions to be made. The struggle over which purpose should be served first led to plans to expedite both channels, with the result that data moved with glacial speed, at best.

The almost constant state of flux in which SQTs moved led to considerable confusion and some frustration in units. In retrospect, it is easy to see why changes were made, as TRADOC and ITED tried valiantly to respond to unit concerns and still retain the validity and reliability of the tests. At the time, however, units had difficulty planning both the testing and their other activities, providing notifications, resourcing the tests, and interpreting the results.

In order to ensure fairness for all Soldiers, the program was astonishingly complex, and the undertaking was enormous for such a tight timeline. The complexity only grew over time, as situations arose that seemed to warrant special treatment or waivers. Both ITED and the proponents were committed to achieving success, but were unable to meet all of the demands for timeliness, high quality, and reasonable costs.

Some of the difficulty in administration and logistics was due to the dependence on printed materials, even for the hands-on tests. The sheer volume of paper overwhelmed ITED and blocked units from being able to efficiently manage the testing.

Perceptions of unfairness in the testing were widespread. Soldiers worried that they would miss an opportunity for promotion because they missed a test window, because their MOS did not have an SQT, because their SQT had been declared invalid, because their job assignment did not include performance of the tasks that would be on their SQT, or because their unit did not take the time to help them practice for the test.

## Recommendations

These recommendations are formulated based on the lessons learned, as described above.

1. Expand the focus beyond measurement of those things that are well defined, such as tasks and task performance. We have successfully demonstrated the ability to measure those more discrete tasks. But the nature of the Army has changed in the last 10 years and measurement of concepts and broad understanding of underlying concepts have become the content areas of the 21st century. We must focus on measuring concepts (e.g., adaptability, leadership) along with the hard skills. When we have established that these soft skills and underlying skills are critical for job success and readiness, then these should also be fair game for measurement. We need to develop methods to measure them.

2. Continue to make efforts to ensure that test developers are trained and qualified. Whether this is accomplished through guidebooks, workshops, or on-line training, it is critical that the test developers have test development skills in addition to military subject matter expertise.

3. Continue to explore and to use novel approaches. Here, technology is our friend. We have at our disposal more tools for analysis, development, delivery, scoring, and feedback than were even dreamed of in 1977. This is not to say that technology will

solve all of the problems. However, we would be foolish to not explore the possibilities.

4. Decide whether the testing is to be for training diagnosis or for personnel management. This decision will go a long way toward resolving how the scoring and the scores will be handled.

5. Make every effort to design a stable system. Change is inevitable, both because the conditions in the Army will change and because what we know about tests and testing will expand. But the initial designs should be both stable and as flexible as possible, with built-in hooks to accommodate revisions, and with the added proviso that changes, when implemented, should be transparent to the user (the test-taker).

   Also related to stability: Do not let the program be too closely associated with one or a few key individuals. This not only increases the risk of sweeping changes when those persons move on, but can also serve to inhibit the gradual changes necessary for growth. Whatever can be done to ensure long-term funding and congressional-level commitment will serve the program well in the long run.

6. Any program with broad testing goals will be complex, and perceptions of unfairness will continue. A realistic understanding of the complexities will help to soften the blow when the conflicting demands for timeliness, high quality, and reasonable costs become apparent. The Army is a much smaller organization than it was in the 1970s and 1980s and individual communication and dissemination of information is well within cost effective means. We need to work those factors to maximum advantage in a revitalized system.

The early SQT advocates had a vision for an Army-wide individual performance measurement system that would reward proficiency, lead to improvements in training and readiness, and move the Army toward a continuously improving organization. The vision came about because the people were permitted to think and innovate, to try out new and different approaches, unconstrained by details or mechanisms of implementation and logistics. Reality necessarily reared its ugly head, however, and the visions, when realized, required considerable negotiation and compromise.

This in no way invalidates the vision, the visionaries, or the continuing need for innovation and high expectations. It only reminds us that visions, once formulated, still need to be tempered by feasibility. The meeting ground is in detailed, comprehensive short-term and long-term plans and commitments that use real world resources not as constraints, but as enablers to bring the vision to reality.

# APPENDIX C
# Review of the U.S. Air Force Promotion Testing Program

**The Air Force Weighted Airman Promotion System (WAPS)**

**HQ, Air Force Personnel Center (AFPC)**
**Air Force Occupation Measurement Squadron (AFOMS)**
**Test Development Flight (AFOMS/TE)**
**Randolph Air Force Base, Texas**
**Visited by PerformM21 project team members February 11, 2003[1]**

**Purpose/Goals of the Testing Program:**

1. The Air Force has a long history of specialty knowledge testing, dating back to the 1950s when it became a separate branch. In the late 1960s, the Air Force Personnel Research Division of Armstrong Laboratory (AFPRL) developed the WAPS, of which promotion testing was an essential part.[2] AFPRL also was responsible to develop the specialty tests. In 1970, the testing branch of AFPRL was transferred to the Air Training Command (ATC) as the forerunner to the AFOMS. Promotion testing in the Air Force has been continuous since the 1960s.

2. The purpose of the Air Force test program is to provide input to the WAPS in determining who should be promoted. Airmen are first tested as Senior Airman (E4) for promotion to Staff Sergeant (SSgt) (E5). Air Force testing continues for promotion to the rank of Chief Master Sergeant (CMSgt) (E9). Airman who are tested are eligible for promotion in that they have met time in service and time in grade requirements and are not ineligible for promotion consideration. Eligibility includes Skill Level qualification.[3] For promotion to E5, E6, and E7, the tests constitute up to 43.5% of the WAPS. For promotion to the grades of E8 and E9, the test constitutes 29% of the weighted (administrative) score and 12.5% of the total score.

3. There is probably some training and readiness benefit realized because of the test preparation on the part of the individual; however this is not a stated goal of the AF test program.

[1] The information contained in this paper is based on that visit and is the interpretation of the authors. This information is not official and does not reflect the position of the Air Force or the policy of AFOMS.
[2] This was based on a mathematical model that identified specific attributes (factors) and the value (weights) used in the existing board-based promotion selection decision process. This was field tested against the board system and found to duplicate the promotion decisions of the boards. Subsequently, the boards for pay grades E5, E6, and E7 were dropped.
[3] AF skill levels pertain to training and are generally, but not precisely, related to pay grade. 1-level (Untrained) designates airmen who are in basic or technical school. 3-level (Apprentice) is awarded after graduation from technical school. 5-level (Craftsman) is awarded after a period of on the job training (OJT) and completion of specified career development courses (CDC). 5-level generally requires, minimally, about 18 month's job experience. 7-level (Supervisor) training usually starts after promotion to E5 and requires more OJT plus 7 level job-school attendance (if available) or other CDCs. 9-level (Manager) is assigned only to E8 and E9 grades.

4. WAPS is used only for promotion of active duty airmen. It is not applicable to the Air National Guard or the Air Force Reserve.

**Test Content:**

1. Airmen who are competing for promotion to SSgt (E5), Technical Sergeant (TSgt) (E6), and Master Sergeant (MSgt) (E7) take two tests. The first is the Specialty Knowledge Test (SKT) that tests the airman in his or her Air Force Specialty (AFS).[4] The second examination is the Promotion Fitness Examination (PFE) that is administered to everyone Air Force wide.

2. Although not all topics are tested, the following is a list of the type of content included on the PFE:

   - Air Force Doctrine
   - Air Force Organization
   - Air Force History
   - Enlisted Heritage
   - The NCO
   - NCO Leadership
   - Military Customs and Courtesies
   - Standards of Conduct
   - Standards of Appearance
   - Enforcing Standards
   - NCO Supervisory Responsibilities
   - Personnel Issues
   - NCO Management Functions
   - Personnel Programs
   - Full Spectrum Threat Response
   - Security
   - Communicating in the Air Force

3. Airmen who are competing for Senior Master Sergeant (CMSgt) (E8) and Command Master Sergeant (CMSgt) (E9) take a single examination called the United States Air Force Supervisory Examination (USAFSE).

4. Test subject matter is based on occupational analysis that is also performed by the AFOMS.[5] Job analysis surveys are conducted in each AFS at least every 3 years and more often if requested or if there have been known changes within the AFS. Each occupational survey requires about 8-10 months to prepare and complete. The current return on surveys is 50-60% and AFOMS has a goal in this area of 75% return.

[4] AFS are organized under Career Fields. There are roughly 150 AFS and 36 Career Fields. Not all AFS have their own test (SKT). If an AFS is exempt from an SKT, the Promotion Fitness Examination is doubled in the WAPS.

[5] The Air Force also has a long history with occupational analysis, going back to computer based systems developed in the 1960s. In 1967, the AF introduced the Comprehensive Occupational Data Analysis Program (CODAP). Occupational analysis and test development programs were combined in the same organization in 1970.

5. Content for the PFE and the USAFSE is determined through a Military Knowledge and Testing System (MKTS) survey. There are 350 general Air Force knowledge topic areas included in the survey; each topic area is included on a survey that is administered to 5,000 NCOs.[6] Respondents are asked to rate the topic area according to the need for professional knowledge or skill based on the rank that the respondent holds. Survey results are compiled with topic areas ranked "high" or "low" based on an average of 5.00 with a SD of 1.00.

6. MKTS results are provided to an MKTS Advisory Council Workshop that meets every other year. The MKTS Advisory Council is chaired by the Chief Master Sergeant of the Air Force (CMSgtAF) and includes all major commands (MAJCOMs)[7] and other selected Chief Master Sergeants. This council determines the general Air Force knowledge and skill areas and levels to include in the Test Preparation Guides along with a determination of which NCO grade should be responsible for what level of knowledge.

7. Based on the outcomes of the MKTS Advisory Council Workshop, the AFOMS produces Study Guides. The Study Guide is published in a single source but in two volumes: Air Force Pamphlet (AFPAM) 36-2241, Volume 1, Promotion Fitness Examination (PFE) Study Guide, and Volume 2, United States Air Force Supervisory Examination (USAFSE) Study Guide.[8] Study Guides are published every 2 years. About 60% of the Study Guide changes each year. Tested items must be found in the current AFPAM 36-2241.

8. The AFOMS also publishes a *WAPS Catalog* annually. This contains the general guidance for study references for the SKT, PFE, and USAFSE. SKT references are specific for each AFS and are generally referenced to a CDC[9]. There is a direct mailing of an initial issue of the pertinent CDC to an Airman whenever he or she first becomes eligible for promotion. There are also direct mailings if a CDC has been revised. But other than this initial distribution and updates, Airmen are expected to maintain their own CDCs.[10]

[6] A minimum of 800 NCOs are surveyed in each pay grade.

[7] There are eight Air Force MAJCOMs.

[8] These are also produced and distributed on CDR. They are also available on the web but web versions are not considered "official." A CDR copy is sent to every Airman eligible for promotion.

[9] CDCs are very detailed, hard copy, study materials. They are a lot like correspondence course materials only all self contained. Most are very good, very comprehensive materials.

[10] About 90,000 CDC are distributed annually. There are some non-CDC references contained in the WAPS Catalog. Each AF unit has a WAPS monitor and the monitor is responsible to ensure that these materials are made available in the unit on a 1:5 ratio. There are also special provisions for handling classified CDCs. About 75% of the SKT are CDC based; the remainder have either specialized Study Guides or are otherwise referenced in the WAPS Catalog.

**Test Design**

1. All three tests (SKT, PFE, and USAFSE) are 100-item multiple-choice examinations.

2. SKT comes in one version. The PFE and the USAFSE are each produced in four versions. The AFOMS produces about 320 different tests annually.

3. All tests are revised annually. Tests undergo either a "major" revision or a "minor" revision. A major revision requires new or rewritten items for 50% of the test content.[11]The major revision requires the participation of a minimum of four subject matter experts (SMEs) and takes about 5 weeks to complete. Each test must undergo a major revision every 2 years. A minor revision is a revalidation and is based on statistical analysis of the test results. A minor revision involves rewriting or new items for a minimum of 10% of the items (the normal experience is 10%-30%). A minor revision requires the services of two SMEs and takes 2-3 weeks to accomplish.

4. Test items for the SKT are based on the Air Force specialty training standard (STS). The STS is an Air Force publication that describes the specialty in terms of tasks and knowledge an airman may be expected to perform or to know on the job. The STS lists both performance tasks and knowledge areas and defines qualitative requirements for each in terms of proficiency levels (for example, Knowledge levels are: Knows Nomenclature; Knows Procedures; Knows Operating Principles; Knows Advanced Theory). The STS does not contain task descriptions but does have reference(s) for each task or knowledge in the task/knowledge listing.

5. SKTs are based on the top 150 or so tasks from the occupational survey that can also be tied to the STS.

**Test Development:**

1. All test development (SKT, PFE, and USAFSE) is centralized at the AFOMS Test Development Flight (AFOMS/TE) at Randolph AFB, Texas. The Flight is made up of approximately 20 military and 35 civilian members permanently assigned. This organization performs all test related activities including management and planning, test construction, test quality control, test production and distribution, and Test Control Officer (TCO) operations.

2. The permanent group in AFOMS/TE is augmented by SMEs for actual test development. SMEs are brought in to Randolph on temporary duty (TDY) for as long as 32 days for test development, although many SMEs are present for much shorter periods. SMEs are NCOs in the minimum pay grade of E7 and represent the AFS for which the SKT is being developed. (PFE and USAFSE have similar groups of SMEs, but at higher grades.) SMEs are organized into AFS test construction teams and each team works with a test management psychologist (of whom there are eight). The psychologists monitor, schedule, and work with different blocks of career fields and also perform final quality control and follow-through on the test. As many as 18-20 test construction teams are on

[11] Typical rewrite on a major revision is around 75% of content.

site at the same time. There are 8-10 test quality control personnel who work with the teams. AFOMS utilizes about 600 SMEs annually in test construction.

3. At least one SKT SME is a person who worked on development or revision of the Air Force Specialty Code (AFSC) CDC.

4. Each test construction group must have a "defendable" test construction process. SKT test items must be linked to tasks from the occupational survey and must be referenced to a CDC or other approved reference. PFE items must be found in the PFE Study Guide.

5. The Air Force does not use any contractors in test construction[12].

**Test Administration:**

1. Test production (printing, binding, packing) is through service contract with the Defense Logistics Agency (DLA) Document Automation and Production Service (DAPS). Overseas shipment is primarily through the U.S. Postal Service (USPS) registered mail and stateside shipment is via Federal Express.

2. Tests are administered during a test window that varies by grade. Test windows vary from about 4 to 45 calendar days.

3. Test distribution is an automated system based on an Oracle database that identifies names, AFSC, and location of eligible airmen.

4. Tests are distributed and administered through the Test Control Officer (TCO) network. TCOs are assigned to Military Personnel Flights (MPF); there are 65 CONUS TCOs and 23 OCONUS TCOs. TCOs for the SKT and the PFE must be pay grade E7 or above and for the USAFSE must be an E9 NCO. Civilian TCOs are GS5 or above. The TCO does not administer every test; TCOs may appoint other Test Examiners for actual test administration. Test Examiners must meet the grade requirements indicated.

5. There are some Air Force units that are designated "geographically separated units" (GSU). These are tested either by the TCO going TDY to the GSU; GSU members being brought TDY to the servicing MPF for testing; or the appointment of a Special TCO by the GSU.

6. Airmen who are deployed are either tested prior to departure on deployment or tested out of cycle upon their return from deployment.

7. The TCO is responsible for ensuring that examinees are eligible, ensuring the correct test is administered to the correct examinees, and for enforcing standardized test conditions. There are established test requirements for noise levels, lighting, space (15 square feet per examinee), temperature, ventilation, and control aisles between examinees.

[12] AFOMS does use some contractor support in their CODAP occupational analysis and has a contractor designing software to support the Centralized Administrative Reporting and Occupational Measurement System (CAROMS) project, which is intended to be a follow-on occupational analysis/survey system.

8. Both the PFE and USAFSE use individual-use test booklets that are sealed until opened by the examinee.

9. Test administration time is set at 1 hour 45 minutes per test.

10. All test related items are designated "Controlled Test Material" and their handling is prescribed at all steps of the distribution and return process. All distribution and return is through TCO channels with only designated persons allowed access. Materials must be stored in safes or vaults with combinations. There are strict test booklet destruction procedures. All answer sheets are accounted for by the TCO and returned by the TCO to the AFPC.

11. TCOs test about 103,000+ airmen annually administering about 200,000+ tests. During their peak test period AFPC receives about 6,000 test returns per day.

**Scoring and Analysis:**

1. Test items are equally weighted one point each. Percent correct scores are reported, which reduces confusion in cases when some of the 100 items on the test are dropped because of poor item statistics. Scores are a total test score; no subscores or task scores are reported.

2. Results are analyzed by test and by grade and by item. Analysis includes applying Item Response Theory (IRT) for scaling individual items and benchmarking across items.

3. Airmen can challenge any item on the test. About 1,500 challenges are received annually.

**Test Security:**

1. Test security is built into all phases of development and administration. The TCO network is specifically set up to enhance and enforce security and access to test materials is strictly controlled particularly at the point of administration.

2. There are specified procedures for reporting potential compromise. Installation commanders are responsible for initiating an investigation of compromise within 24 hours of any report.

3. Much of the focus of test security is on the individual airman. Violations are subject to prosecution under the UCMJ. The following are examples of specifically prohibited activities:

    - Reviewing or having access to actual test material.
    - Reviewing or having access to illegal study materials that reveal the content of actual or suspected test material.
    - Questioning examinees for the purpose of determining test content.
    - Reproducing or copying any test material.

- Training that concentrates on "teaching the test" or that emphasizes information known or believed to be part of a test.
- Reviewing contents of tests by inspection team members or any other reviewing officials at any level of command.
- Taking a test and claiming to be another examinee.
- Opening a package marked "Controlled Test Material" unless specifically designated to receive and open such materials.
- Improperly storing test materials.
- Gaining access to any security container containing "Controlled Test Material" unless specifically authorized.
- Discussing actual test material or the specific contents of testable material in such a way to highlight actual or potential test material.
- Removing test material from the testing room.
- Using highlighted or marked testable material that reflects actual or believed test material that is shared between, used by, or observed by more than one potential examinee.
- Using marked pretest that are shared between, used by, or observed by more than one potential examinee.
- Promotion tests are designed to measure an *individual's* knowledge of test material, therefore, group study is not permitted.

**Interfacing With Candidates:**

1. Preparing for the test is an individual responsibility. The testing window for pay grades E6 and E7 begins (usually) in February and the window for pay grade E5 opens in May. The Air Force Personnel Data System identifies the eligible airmen in the previous August. The WAPS Catalog (showing the test references) is also published in August. E6 and E7 skill level CDCs are mailed to individuals in September and E5 pay grade CDCs are individually mailed in October. Distribution of the Study Guide for the PFE and USAFSE is on a similar schedule.

2. As noted in the section on test security, preparation is very much an individual requirement. Unit sponsored preparation for the test *per se* is not conducted.

3. Individuals are given post-test feedback as to their percent correct on the test. A single score is provided.

**Associated Policies:**

1. The Air Force does not use promotion boards for promotion to pay grades E5, E6, or E7. Promotion is based on six factors:

   - Decorations = 25 points
   - Time in service = 40 points
   - Time in grade = 60 points
   - PFE = 100 points

- SKT = 100 points
- Enlisted Performance Report (EPR) = 135 points[13]

2. The Air Force uses a centralized promotion board for promotion to pay grades E8 and E9. Board review is limited to a records review (there are no personal appearances). Promotion criteria for these grades are:

    - Decorations = 25 points
    - Time in service = 25 points
    - Time in grade = 60 points
    - USAFSE = 100 points
    - EPRs = 135 points
    - Board = 450 points

3. In the Air Force, promotion vacancies are not based on job vacancies in the AFS. Promotions are made by AFS but they are based on the overall vacancy spread over all AFSs. For example, if the annual vacancy in E5 authorizations is determined to be 25%, then the top 25% in each AFS will be promoted.[14] Additionally, the Air Force can add 5% promotions to those AFSs that are designated as "critically-manned."

4. There is another Air Force promotion program called Stripes for Exceptional Performance (STEP) in which "truly outstanding" airmen can be promoted to pay grades E5, E6, or E7 without regards to promotion standing or to quotas. This is a limited program, usually only about 2-300 in each grade annually.

5. All annual Air Force promotions for a given grade take place in a given month; that is, they are not spread out over the year.

6. The Air Force has current active duty enlisted strength of about 289,000 airmen. They roughly break out as follows:

    - E1-E3 = 80,000 = 27.5% of force
    - Senior Airman (E4) = 53,000 = 18.5% of force
    - SSgt (E5) = 72,000 = 25% of force
    - TSgt (E6) = 44,000 = 15% of force
    - MSgt (E7) = 31,000 = 10% of force
    - SMSgt (E8) = 5,500 = 2% of force
    - CMSgt (E9) = 3,000 = 1% of force

[13] The EPR rates each airman on conduct, duty performance, leadership abilities, dress and appearance, and communicative abilities. Airmen are given a numerical Promotion Recommendation rating from 1 to 5 (top) in these areas. (There is also a narrative portion to the EPR.) Five years worth of reports are consolidated based on a time-weighting factor with the most recent report being worth 10 times the oldest report.

[14] Fractions are rounded up. Since there are many low density AFS, this usually results in higher than authorized numbers being promoted. This can also lead to some seeming AFS differences. For example, promoting 25% of an AFS with 10 eligibles in it is actually a promotion rate of 30%

7. Air Force promotions tend to be slower than in the other services. The following time in service averages for promotion are based on historical data:

    - To E4 = 36 months TIS
    - To E5 = 54 months TIS
    - To E6 = 156 months TIS
    - To E7 = 204 months TIS
    - To E8 = 240 months TIS
    - To E9 = 265 months TIS

**Links to Other Air Force Systems:**

The testing system is integrally aligned with the Occupational Analysis and with Professional Study Guide development, all falling under the AOFMS organizational umbrella. AFOMS is an operating entity under the Air Education and Training Command, Director of Operations, which is a General Officer (GO) slot.

**Self Assessment:**

1. There are a number of official sources for assessing one's overall promotion eligibility and standing. There are no official self-assessment instruments (practice tests) for either the PFE or the SKT.

2. There are a number of commercial available practice tests advertised on the web, but none are endorsed by the Air Force and government funds may not be used to purchase them.

**Test System Organizational and Policy Structure:**

Test development, printing, ordering, and shipping activities are centrally located at AFOMS. Responsibility for test administration, scoring, and promotions is centrally located at the Air Force Personnel Center (AFPC), which is a field operating agency of Air Staff. AFOMS and AFPC work very closely in the conduct of the promotion program.

**Program Evaluation:**

1. All test program annual and recurring evaluation is conducted internal to the AFOMS. AFOMS is, of course, answerable to its higher headquarters (AETC) in this regard.

2. The Air Force had not had an independent review of its test program since 1969. In 2002, the Air Force commissioned Chauncey Group International to conduct an audit of the testing portion of WAPS. A CGI group of industrial psychologists and psychometricians visited with Air Staff, TCO, and AFOMS in a validation of Air Force test policies for

compliance with test development and administration standards and a verification of compliance with those policies. Overall, the program was given a very positive review[15].

## Other Observations/Comments

The Air Force does not have a current plan to go to an automated testing system or to use computer administered tests. There are two primary, related reasons: First, there is no existing infrastructure within the Air Force to support such a testing system. Given no existing infrastructure, the estimate is that administration costs would increase by two to three times.

[15] The 11 evaluation areas were: Customer Service, Fairness, Uses and Protection of Information, Validity, Assessment Development, Reliability, Cut scores, Scaling, and Equating, Assessment Administration, Score Reporting, Assessment Use, Test Taker's Rights and Responsibilities.

# APPENDIX D
# Review of the U.S. Navy Promotion Testing System

**The Navy Enlisted Advancement System (NEAS)**

**Navy Advancement Center (NAC)**
**Naval Education and Training Professional Development Center (NETPDTC)**
**Pensacola, Florida**
**Visited by PerformM21 project team members February 3, 2003[1]**

## Purpose/Goals of the Testing Program:

1. The purpose of the Navy test program is to rank order all qualified candidates on the basis of technical and general military knowledge for promotion to the next higher pay grade. All candidates who take the promotion (advancement) tests are already qualified to advance by other factors (time in pay grade, commanding officer's recommendation). The test is therefore designed to spread out the pool so the top test performers can be compared to others who took the same test. It is not the only factor that is quantified in Navy enlisted promotions, but it accounts for between 30% and 60% (depending on rating) of the sailor's promotion score (called the Final Multiple Score [FMS]).

2. There is probably some secondary effect in training and readiness because the tests are directly linked to job standards, and study references (the bibliography or "bib") are provided to candidates. Presumably, because of the study references, there is some element of self-development involved. However, the main effect is that sailors that are top performers in terms of relative test scores are the ones promoted. The Navy maintains that test performance is based on the knowledge and experience gained in the sailor's years in the Navy and not on a few weeks cramming for the examination.

## Test Content:

1. Test content must be basically "validated" against Occupational Standards (OCCSTDS) and Naval Standards (NAVSTDS), which are published standards for all Navy ratings[2]. OCCSTDS apply to the job (rating) (such as Boatswain's Mate), while NAVSTDS are applied to all personnel in a specified pay grade. OCCSTDS tend to be task-based while NAVSTDS tend to be knowledge-based. Examples of NAVSTDS categories are Naval organization, leadership, programs and policies, professional development, training, personnel safety and damage control, CBR, security. These are also known as "professional military knowledge" (PMK) areas.

[1] The contents of this paper are the results of that visit. Information contained in this paper is the interpretation of the authors and does not reflect official policy of the U.S. Navy or of the Navy Advancement Center.

[2] "Ratings" are broad enlisted career fields that identify occupational specialties that have related aptitudes, training, experience, knowledge, and skills. Ratings are of three types: general, service, and emergency (wartime). There are about 85 ratings, although there are subdivisions and specialties within the ratings. Ratings are organized into Occupational Fields, of which there are currently 23.

2. OCCSTDS and NAVSTDS are both based on periodic (supposed to be updated every 3 years) fleet surveys of incumbents and operational community representatives. These surveys are conducted by the Naval Manpower Analysis Center (NAVMAC) in Millington, TN and are not directly aligned with NETPDTC. The Navy is moving toward a revised job analysis process (involving identification of "skills objects" and "skill tasks" which, when approved, will be used as part of the validation of future tests.

3. Tests for sets of ratings are developed in a team with subject matter experts (SMEs) and headed by a civilian test psychologist. There is usually one SME per occupational area. SMEs are Chief Petty Officers (CPO) in grades E7 thru E9 who are brought into NAC usually for a 3-year tour. SMEs largely come to NAC directly from a fleet assignment. They are responsible for updating job performance data based on their experience. There must be an OCCSTDS or NAVSTDS or other standard to support the content and every question must have a "bib" reference.

4. There are separate teams that prepare the occupational tests and the PMK tests. Occupational teams integrate the designated PMK items as a block into PMK sections of the test. Occupational test developers do not do any content selection or test development of PMK items.

5. Tests are 200 items and, by policy, the content is divided between Occupational Items and PMK Items as follows:

| Grade Tested | Occupational Items | PMK Items |
|---|---|---|
| E7 | 100 | 100 |
| E6 | 115 | 85 |
| E5 | 135 | 65 |
| E4 | 150 | 50 |

6. The occupational SME picks the content areas for testing based on his/her own experiences for priority or criticality; not all the content areas in a rating will be on any single test. However, there is an effort, over time, to cover all the content in a rating. The SME and the test psychologist prepare a test blueprint (which the Navy developers also call a "test plan" and a "test outline") based on the SME's fleet experience and input. In reality however, because of the long testing experience, most blueprints are updates of the previous blueprint rather than completely new blueprints. (NOTE: PMK content areas are fairly new [2002] so there is not as much history with them.)

7. By policy, the minimum number of sections (content areas) for each pay grade is six (combination of occupational and PMK). The maximum number of sections differs by pay grade:

   E4 – 15
   E5 – 13
   E6 – 11
   E7 – 10

The minimum number of items (questions) per section is 10 and the maximum number is 25. Collectively, tests must total 200 items.

8. Tests are developed for each pay grade of each rating. As long as content is supported by OCCSTDS/NAVSTDS and has supporting bibs (references) it is fair game to be included in the test. There is no attempt to track for specialized equipment or duties. Everyone in a rating and pay grade (collectively called a "rate") takes the same test, so everyone ends up getting test items that may not be applicable to their location or training.

9. The PMK items are identical for all sailors, across ratings. The content changes as the pay grade changes.

**Test Design:**

1. The NAC develops about 575 tests annually (E4, E5, E6 sailors are tested twice each year x approximately 85 ratings = 510 tests; E7 sailors are tested annually x approximately 65 ratings = 65 tests). By policy, each test must have approximately 40% "new" items minimum. The remaining 60% can come from the item bank and are called "control items." There are strict rotation rules on the control items. (Because PMK is so new [2002], there are currently relatively few PMK control items.)

2. The goal of the test is to spread people out for promotion purposes, so the ideal test is one in which one-half the test population scores at the 50th percentile. On items, they aim for $p$ values of .30 to .60 (mean target = .54).

3. As mentioned, all tests are 200 items – multiple-choice, four response option, single correct. About 10-15% of the tests use figures or illustrations; most are purely text-based. The goal is to make the occupational items as performance-oriented as possible. Depending on the test they will be given, candidates may be told to bring single line (non-programmable) calculators, slide rules, bearing rate computers, nautical slide rules, speed solvers, towed array range finders, musical manuscripts, drafting equipment, maneuver boards, or other job-specific pieces of equipment to solve performance problems.

4. All tests are paper-based with answers recorded on a scannable answer form. Tests are designed to be completed in a 3-hour test time, within a 4-hour test block administration window.

5. Tests are prepared in two formats: A and B. The A format is the test that almost everyone takes. The B format is called the "substitute examination" and is a previously used form of the exam. Substitute examinations are used for candidates who did not take the examination on the regularly scheduled day.

**Test Development:**

1. SMEs (also called exam writers) are E7/E8/E9 sailors brought in for a minimum 2-year assignment to NAC. The selection process (to the extent there is one) varies by rating.

SMEs go through a 4-day training and orientation period and, situations permitting, overlap with their predecessors. There are about 85 SMEs, covering all occupational areas, organized into 13 teams. Each team works with a civilian test psychologist who also acts as a reviewer/editor. There are three groups of test writing teams, the largest containing all unclassified ratings, another containing classified ratings up to Secret, and the last team (headed by a permanently assigned CWO) focuses on about seven Cryptographic ratings.

2. Except for information technology (IT) support from DynCorp, the Navy does not use any contractors in its test development.

3. Exam writers follow a procedure in an in-house publication called The Item Writer's Standards Guide that contains the rules of item writing. Following the blueprint, item writers are free to develop any item that they want. However the item (answer, procedure) must be supported by a doctrinal publication that must be referenced and that citation becomes part of the record supporting each item. That same reference becomes part of the "bib" which is made available to the candidates so it must be something that is accessible to them.

4. Exam writers can select approximately 60% control items – items with little or no modification that have previously been used. Previously used items that are improved through more major revisions based on item analysis diagnostic information are also allowed.

5. Tests and test items are submitted for peer and team reviews during development to look for wording consistencies, trivia, trick questions, excessively hard or excessively easy items, reasonableness of alternatives, and comparability of alternatives. Each exam is read aloud in its entirety during a team review.

6. NAC uses a computer system called "examination development software" (EDS) in its test development. This operates on a closed LAN system only at the NAC location in Pensacola. EDS is supported by large databases that contain past examinations, bibs, references, item banks, graphics, evaluations from test administrations, and OCCSTDS, to name some of the known databases. Item development and test section construction is done directly on EDS.

7. There are no pilot tests administered, though the test development SME "takes" the test as quality control check. Adjustments are made to the test post-administration as necessary. Item analysis from past tests is used by SMEs to review and cull items that may have deficiencies.

**Test Administration:**

1. Tests are administered twice a year (in March and September) for promotion to E4, E5, and E6 and once a year (January) for promotion to E7. To be eligible, sailors must meet certain statutory requirements (such as time in pay grade) and have their commanding officer's recommendation to be promoted.

2. The test window is 3 hours on the same day, worldwide. A substitute test is allowed at a later date for those who miss the test window, but "misses" are very tightly controlled and this is not treated as a make-up date.

3. Test candidates are identified by Time in Rate (TIR) by NETPDTC approximately 3 months prior to testing and lists are posted on their Internet site for review by unit exam representatives (generally Education Services Officers [ESOs]), who alter the lists as needed. One and a half months prior to the exam date, tests and answer sheets are mailed to the command UIC for the sailor. Test answer sheets are sent out already bar-coded with identifying information and contain header information already completed including information on other promotion factors.

4. Tests are sent to the ESO who is responsible for administering the test. The ESO appoints proctors (who must by CPOs or above and outrank the sailors being tested) to assist him/her. ESOs are responsible for training the proctors. There must be a minimum of 1 proctor for each 25 examinees.

5. ESOs are provided an instruction booklet for each exam administration. This booklet contains four parts:

   Description of the Answer Sheet
   Instructions to the ESO Prior to Administration
   Verbatim Instructions to the Candidates
   Post-Administration Instructions

6. In 2001, NAC shipped 371,430 regular active duty exams and 70,786 Naval reserve exams. (This is all ratings and pay grades E4 through E7.) Additionally they shipped 27,278 Substitute Exams. In return, they processed 242,998 active duty answer sheets and 42,099 reserve answer sheets. The introduction of the website eligibility notice and bar-coding just started in 2002 and NAC is confident that this will reduce the discrepancy between tests shipped and tests processed.

7. Test production (printing, binding, packing) is though a service contract with the Defense Logistics Agency Document Automation and Productions Service (DAPS). Shipment is by UPS, FedEx, USPS Express Mail, and USPS Registered Mail (for classified tests). FedEx is the current GSA express mail contractor and is used for answer sheet returns where possible. However, UPS and FedEx cannot be used for direct shipments to APO/FPO addresses.

8. The Navy's cost for printing, labor, and packaging is $2.25 per exam. Shipping per exam is between 20 and 40 cents, depending upon mail carrier.

**Scoring and Analysis:**

1. Item analysis starts as soon as examination results start feeding back on examination day. Occupational items are analyzed by rating and pay grade and the results sent to the occupational exam writing team. PMK items are analyzed Navy-wide by pay grade and sent to the PMK exam team. A statistical evaluation report (STAR) is prepared to identify

problematic items. Items determined to be unacceptable are not included in the final score.

2. Only correct answers (out of the 200 items) are credited and there is no penalty for wrong answers (guessing). There are no adjustments for unanswered items.

3. The raw scores are converted to standard scores (for the overall score) and norm-referenced percentile scores (for section scores) for reporting purposes.

4. Analyses are conducted to estimate test reliability (internal consistency estimates) and to compare exam and exam section difficulty to previous years' exams.

5. Subgroup "verification" analyses (gender, ethnicity) are conducted. It is not clear from our available information exactly what types of analyses are conducted.

6. NAC conducts a number of statistical analyses to identify evidence of test compromise. These analyses focus on candidates scoring at or above the 98th percentile and candidate pairs that have an inordinate number of identical incorrect responses.

**Test Security:**

1. Test security is built into all phases of development and administration. The EDS is a closed LAN with password and accessibility restrictions. Administration is tightly controlled and monitored through the ESO system. All tests and answer sheets are controlled items. Positive controls and individual identification verification are used during test administration. Proctors are present at a 1:25 minimum.

2. NAC says they have never lost a shipment of tests because of the effective controls applied by Federal Express and registered mail. They have had some return answer sheet losses because commands sometimes fail to return answer sheets by traceable means.

3. NAC employs a well-developed process of statistical detection of examination compromise. Sometimes this is done by programmed computer searches and sometimes in reaction to anonymous tips or to discrepancies reported by ESOs. NAC has established statistical documentation of cheating in several instances in the past and this statistical analysis has been accepted and sustained in courts martial.

4. Investigations of suspected compromise are conducted by the Office of Naval Investigations (ONI), the criminal investigation arm of the Navy. Disposition of compromise cases is made by the chain of command, however they are aggressively pursued. At the very least, suspect candidates are told to retest.

**Interfacing With Candidates:**

1. It is up to the individual sailor to prepare for the examination. The bibs for each test are available electronically on the NAC website (https://www.advancement.cnet.navy) or they are available as a supply item via special channels for locations that cannot download from the website. The bibs only list the references – it is still the sailors'

responsibility of obtain those references or pursue other methods of study. NAC is also making available Advancement Strategy Guides online. Sailors are *not* given a copy of the test plan or outline.

2. The bibs are posted approximately 5 months in advance for each of the E4/5/6 tests and 6 months in advance for the E7 test.

3. Group study for tests is allowed, however there is little first hand information on how commands actually apply this.

4. Each sailor gets a post-exam profile report. It shows the candidate's overall standard score and the average standard score across all candidates. For each section of the exam, it also identifies the number of questions in the section, how many questions the sailor got correct, (raw) and then the percentile ranking for each section. The profile does not identify which questions were answered correctly.

5. Despite providing profile information, the Navy warns about using the profile as a basis for preparing for the next examination. They maintain that each examination is unique and stand alone and there is no carryover from the exam just completed.

6. The profiles and feedback are not immediately available to sailors. Because tests are norm-referenced, all test results must be accumulated and statistical anomalies or disputes resolved before final profiles can be provided.

7. The Navy has an active program to inform sailors about the testing program and the advancement program in general. They have a "Brief to the Fleet" which is a face-to-face program delivered by a senior CPO, which last year had small group contact with 3,700 sailors. There is also a website (see above) that contains a lot of specific information including a FAQs section about promotion and testing. The Navy realizes that there is a problem getting accurate testing/advancement information to the fleet.

8. Even with these outreach efforts, the Navy responds to thousands of inquiries about the testing program from sailors each year.

**Associated Policies:**

1. Promotion eligibility is determined by time in rate, schooling (in some cases), and the commander's recommendation to promote.

2. Sailors accumulate a Final Multiple Score (FMS) that determines if they will get promoted. For pay grades E4/E5/E6 there are five factors: (For E7 sailors, there are only 2 factors: Test results and Performance.)

   Test results
   Performance (Evaluations/FITREPS)
   Service in Pay grade
   Awards and Decorations
   Passed-not-advanced (PNA) Score

3. The test results are weighted differently in FMS by pay grade. For pay grades E4/E5, it counts 34%; for pay grade E6, it is 30%; for pay grade E7 it is 60%.

4. Promotions are still made by vacancy in pay grade by rating. Cut scores are established at Department of the Navy level for each promotion period based on those vacancies. The Navy has two promotion periods to pay grades E4 through E7. E4 through E6 selectees are advanced over a 6-month period on the 16th of each month. For promotion to E7, the FMS is used to establish Selection Board Eligibility, after which an OPNAV Selection Board makes the decision. E7 selectees are advanced over a one-year period on the 16th of each month.

5. Sailors in pay grades E4 through E6 who are not promoted can accumulate points from their last test to carry over in their FMS. This is the PNA (passed not advanced). The PNA points are awarded based on the sailor's examination standard score from the last test combined with a part of the fitness rating. These are recalculated with each semi-annual examination cycle and keep accumulating up to a total maximum of 30 extra FMS points. So the program eventually rewards sailors who do well on every test, even when ratings have high cutoff points.

6. Candidates can challenge information in the test either at time of testing or as follow up. However most exam discrepancies address header information. Internal quality control and statistical review is the main source of content review.

7. Profile sheets (test results) are provided to commanders via the web so they can congratulate successful sailors.

**Links to other Navy Systems:**

The NAC tests are administered to the United States Navel Reserve (USNR), although on a different schedule.

**Self Assessment:**

There are no separate versions of the actual exams offered as practice tests. (Note, however, that there are practice tests and test questions available commercially on the web and from other sources for a price. NAC maintains that these are worthless.)

**Test System Organizational and Policy Structure:**

The Navy system for development, administration, and policy is centralized at the NAC.

**Program Evaluation:**

Most program evaluation is conducted internally to NAC both for individual tests and the test program overall.

## Other Observations/Comments:

1. They would like to add skill-based (hands-on) tests to the program. The Navy personnel research group is looking into this (Task Force Excel).

2. Because most E4s get promoted, the test is not very useful for supporting that process.

3. The Navy is starting to use SkillSoft to collect job task analysis data that they hope to start using to support test plan development.

### Update – Computer Testing

Subsequent to the date of this visit, and during a follow up conference with NAC in November 2003, we learned that the Navy is actively exploring the feasibility of administering computer-based tests. Over the summer, a group did considerable research and study on various programs and have a pilot program ready to try out in early 2004. The key features of the trial program are as follows:

- The test administration software is Perception. This was chosen after a review of a number of different testing software.
- Administration will be via a combination of laptop computers sent to the fleet and web-based access for shore activities. Examinees using laptops will take the test on disks (or Secure Digital [SD] cards) that will be handled separately from the computers; no tests will be installed on the computer hard drives. This addresses certain security and storage issues.
- The Navy has modified its EDS test development system so that it can export a Perception-compatible question file.
- The emphasis has been on taking advantage of the computer's multimedia capability, not just to provide an electronic version of the paper-based exam, but to also increase the capability of assessing a sailor's ability to perform required tasks.

# APPENDIX E
# PerformM21 Technology Requirements

*Prepared by David Katkowski, Stephanie Itchkawich, & Jeffrey Barnes*

This document provides information and recommendations about several important considerations for developing and administering an individual performance assessment system for enlisted Soldiers:

- Test development/delivery software
- Test delivery portal – Digital Training Facilities (DTFs)
- Web-basing requirements
- Administrative support

Though we have learned a great deal about what these considerations entail, there is more work to accomplish. We will continue our efforts at gathering pertinent information until we are able to settle currently unresolved issues and delineate a solid course of action.

## Test Development/Delivery Software

### *Background*

In the early 1990s, developmental and administrative burdens forced the Department of the Army to halt its administration of the SQT. Since then, developments in computer-based testing software have provided an economical vehicle for the administration and maintenance of large-scale testing programs, such as the SQT. Intranet and Internet test administration, user-friendly test development tools, "spot-on" item feedback, and instantaneous report generation are all features that have substantially reduced costs and increased efficiency. The Department of the Army will have to make use of these developments if it is to realize an operational and affordable individual performance assessment system for enlisted Soldiers; therefore, we searched for a test development/delivery software package that would suit the Department of the Army's needs.

To ensure a proper fit, we evaluated several test development/delivery software packages. First, we conducted an exhaustive Internet search and examined each of the results. Then, we reduced the field to four promising solutions and evaluated them extensively: FastTEST Pro®, BlackBoard®, SMT-PCTest/SMT-Bank®, and Questionmark's Perception®. Though all of these packages had advantages and disadvantages, Perception provided the most robust, customizable, and affordable solution. At the time we undertook this review, it was evident that Perception had had the opportunity to mature over more than 10 years of development compared to only a few for the other software packages. In addition, the Navy conducted its own extensive evaluation of test development/delivery software packages and determined that Perception was the most suitable for its annual enlisted testing program.

## Description of Perception

Perception provides everything needed to author, administer, deliver, and report on computerized assessments. The software package has three components that interact with each other to accomplish these ends: Question Manager, Assessment Manager, and Enterprise/Windows Reporter.

Perception's authoring system allows test developers to create, modify, and delete questions and assessments. One can author in a Windows-based environment, via a Window's-based program, or in a Browser-based environment, via a server-based system. The Window's-based authoring requires the installation of software on a PC and provides a powerful environment for creating simple or very complex assessments. Browser-based authoring provides a simpler environment, and allows wide deployment of authoring features via a browser. Both authoring environments allow banking of items with Perception's Question Manager module and building of assessments using Perceptions' Assessment Manager module.

Perception's administration system allows maintenance of information about participants, assessment schedules, groups, administrators and their privileges, web servers, and Perception's e-mail broadcast system. Administrators can handle everything within Perception's framework.

Perception offers a variety of delivery options. It can be used on the Internet, Intranets, and individual Windows machines or networks. Another delivery option is Perception to Go, which allows participants to synchronize and then go offline while they complete assessments. Perception also makes it possible to print assessments from the Windows authoring system for distribution on paper.

When examinees respond to items, Perception writes the information to the answer database as soon as the assessment is started and it updates that information as the participant progresses through the assessment. One can provide real-time reporting on results while candidates are taking assessments or wait until they are finished. Perception's pre-defined reports provide comprehensive statistics to assist in analyzing the results of an assessment or group of assessments. Results can be presented in a variety of formats to provide different forms of analysis. Perception comes with a set of report templates and allows one to customize reports. Results are stored in MS Access by default, but can also be stored in MS SQL or Oracle.

## Perception's Advantages

Perception offers several advantages with its user-friendly interface, integration capability, added security, online documentation and knowledge database, technical support, and customizability. Each contributes to the cost effectiveness, maintenance, and deployment of large-scale testing programs. With Perception's authoring software, anyone can write items and develop assessments. There is no need for trained programmers constantly updating and changing assessments. The people who write the content of the items can also take responsibility for computerizing it.

Perception also provides numerous ways to integrate with its delivery system. This enables third party management systems to launch and track assessments. There are three industry standards available to launch and track assessments:

- **AICC**
  The Aviation Industry CBT (Computer-Based Training) Committee (AICC) is an international association of technology-based training professionals. The AICC develops guidelines for aviation industry in the development, delivery, and evaluation of CBT and related training technologies.

- **SCORM**
  The Sharable Content Object Reference Model (SCORM), produced by the Advanced Distributed Learning organization (ADL), allows the sharing of content through the use of a defined data model and runtime environment.

- **Web services**
  Questionmark Web Integrated Services Environment, QMWISe, is a comprehensive series of web services methods that act as an Application Program Interfaces (APIs) that enable registration, management, and reporting systems to tightly integrate with Questionmark Perception.

In addition to its delivery system integration capabilities, Perception can integrate items programmed in Java or Flash, graphics, and video clips. This allows the item writer to develop more realistic and applied test items than traditional multiple-choice tests. This can lead to a better measure of examinees' knowledge, skills, abilities, and other characteristics. It also reduces the reading requirements, which might reduce the risk of adverse impact.

A utility called Questionmark Secure adds needed security while delivering assessments using a web browser. This feature allows one to prevent examinees from printing questions, using the right-click on the mouse, saving the HTML, viewing the source, opening new applications, task switching, and accidentally exiting an assessment in a proctored environment. While this utility does help reduce the risk of different forms of cheating, it is not a substitute for a proctored testing environment.

Perception's online documentation and knowledge database are large and up to date. It also provides access to white papers, tutorials, web seminars, and FAQs on a wide variety of topics. In addition, Quesitonmark sends e-mail updates on recent developments, future directions, and solutions to common problems.

### *Perception's Limitations*

Most of Perception's limitations are small relative to its strengths and there are easy fixes that should be included in future versions or could be remedied by Perception's consultants in a customization partnership. However, the present lack of certain functions can cause undo frustration on the part of the user. A savvy user may be able to "add" these functions to the databases that they have access to, but this can be risky and cause more harm than good.

First, Perception does not allow a simple count function. Obtaining a count of the number of items by blueprint category is often desirable by item developers for content balance. It seems the only way to obtain the desired counts is to manually count the number of items in each category. With large question databases, this is extremely time consuming and susceptible to human error.

Perception also assigns a unique ID number to each item, but does not allow the user to view it easily in the question database. Instead, the user must assign a unique ID number to each item. This is not desirable because errors could easily result in assigning the same ID number to multiple items, especially with large question databases.

Finally, Perception allows users to enter notes in a notes field for each item; however, users can only view notes within each individual item. This is acceptable with small question databases, but can become time consuming and counterproductive with large ones. In addition, Perception does not allow users to print notes with items. Typically, users include item references or reviewer recommended changes in the notes field. This is important information that subject matter experts (SMEs) require access to when reviewing items. Similarly, Perception does not allow the printing of the database directory by item. After the SMEs review a set of printed items and suggest revisions, the user must navigate to the items in the question database to make the necessary revisions. This is an extremely difficult task without a specified database directory for each item.

*Summary*

In summary, the Department of the Army should make use of a modern test development/delivery software package to minimize the administrative and developmental burdens associated with past large-scale testing programs such as the SQT. Based on our evaluation of such software packages, we feel that Questionmark Perception offers the best overall utility for the administration and delivery of an operational and affordable individual performance assessment system for enlisted Soldiers, at least at the present time.

## Test Delivery Portal – Digital Training Facilities (DTF)

### Background

The Army's Distributed Learning System (DLS) delivers training and education programs to military and civilian personnel at hundreds of Digital Training Facilities (DTF) worldwide. DLS, formerly The Army Distance Learning Program (TADLP), got started in the wake of the 1997 Quadrennial Review. The report gave rise to a Department of Defense-wide strategy to use information technologies to modernize education and training. The Army DLS covers active-duty personnel and the Army Reserve, in addition to civilians. The National Guard runs a sister program called the Distributed Training Technology Project (DTTP). DLS is supported by the Program Executive Office Enterprise Information Systems (PEO EIS) and by the U.S. Army Training and Doctrine Command (TRADOC) Program Integration Office. The

DLS has a large geographic footprint, carried out in coordination with the Army Network Enterprise Technology Command (NETCOM) and PEO EIS.

DLS manages all of the DTFs from the Enterprise Management Center in Fort Eustis, VA. A central manager can release virus updates and operating system updates, carry out bandwidth analysis, and give each student individual user IDs and passwords from an enterprise level. This allows the Department of the Army to enforce standardization across the facilities, but lifts the burden off the individual IT managers.

DLS is an unclassified project, but security is still critical because of its significant presence on the Army's network. To prevent the possibility of any hacker accessing a classified system via the DLS, system designers put particular emphasis on user IDs, passwords, and other security measures.

DLS is rolling out the program based on a "block" strategy. Block I involves acquiring the computers, video equipment, and other hardware necessary for stand-alone computer-based training. Block II entails networking those facilities in order to allow access to the Internet for Web-based training, chat rooms, instant messaging and the like. Block III is the integration with Army training management systems, while Block IV consists of the deployable digital training facilities.

Complete implementation of the Army DLS is slated for 2010. It involves fielding DTFs throughout the world, as well as deployable training facilities for support of the field Army. TRADOC has fielded seven prototype deployable training facilities and two more are in development. The goal is to ultimately provide 95% force coverage. Over 200 DTFs are operational in the United States and at installations in Germany, Belgium, Italy, Korea, Japan, and Okinawa.

## Description

The typical DTF on an active post is equipped with 16 computer systems – eight tables each housing two computer stations. Each computer is capable of connecting to the Internet as well as the digital training network. A computer station is comprised of an Intel Pentium 2 or higher processor, CD-ROM optical drive, 17" monitor, standard U.S. keyboard and mouse, running on Windows 2000. Other resources include a network printer, manager's workstation, and a local server. Some DTFs are also equipped for tele-training, with a large video monitor at the front showing the remote classroom. A camera mounted above the monitor allows a remote instructor to see the class.

While some degree of variation from facility to facility might require localized adjustments, the facility we viewed at Fort Myers displayed most of the core criterion desired.

- The room was appropriately sized, well lit, and comfortable, admitting no through traffic.
- Sufficient space existed to allow proctors to move freely about during an assessment without disrupting test takers.

- No risers or steps were present which would assist a participant in viewing materials displayed on another participant's monitor.
- Desks were constructed in such a way as to prevent participants from easily accessing ports on the back of CPUs.
- Front plates of CPUs and peripheral keyboards did not house USB or other ports to which data storage devices might be attached.
- CPUs were Intel Pentium II or higher; standard plug-in features were available.
- Facility managers possessed the ability to suppress printing from workstations and printers were physically distant from workstations.
- Facility managers appeared to be helpful and knowledgeable about the testing process and alert to possible sources of compromise.
- Facilities were represented as being convenient to nearly all personnel, accessible by joint service members, and available to National Guard and Reserve components.
- Facilities were represented as being utilized significantly below current capacity.

Currently DTFs are in two configurations as follows:

- BLOCK 1 Configuration: Those facilities that have not been configured to interact with the DLS network. These sites have workstations but do not have access to either the Internet or communication with other systems.

- BLOCK 2 Configuration: Those facilities that have been configured to interface with the DLS network. Workstations at these sites have the capability to communicate with the DLS network and its various resources. These include access to the Internet and collaboration with other DLS workstations.

Although personnel testing was not included as a DTF mission, these facilities provide all of the necessary components to successfully implement computer-based testing. The "good news" for DCAP is the groundwork has already been established for Internet testing at DTFs by the NCO Leader Skills Inventory (NLSI).

### Areas Potentially Requiring Accommodation

Areas where accommodations may require adjustment or may need to be addressed are:

- Full access to the Internet is currently permitted, allowing for the potential of opening windows with test cribs, resource materials, or uploading assessment items.

  This may be addressed through the use of secure browsers during assessments. A browser of this nature is supplied with the package that the assessments are being designed and delivered in. The browser for the local user is free and requires 5MB or less of space to deploy on the user's workstation. We are informed that this would require a review of the software by Army computer security professionals. It is not anticipated that this application would generate concerns upon such a review. There is an additional expense involved in acquiring the server component that interacts with the secure browser during delivery.

- A lack of partitioning between workstations.

  This may be addressed through the employment or retrofit of temporary partitions. This may be overlooked if assessments are designed which scramble both the order in which test items are presented and the ordering of the distractor answers presented for each question. In this manner, the chances that a participant may benefit from opportunistically viewing a neighbor's response are drastically reduced.

- Formalizing the protocol for clearing workstations before and after assessment administration. Currently procedures are in place to conduct activities of this nature; however, these protocols may need to be revisited to ensure that there are no new concerns generated by this assessment process.

- Creating a protocol for delivering assessments that may involve sensitive data or procedures. For MOSs that may involve such data: dedicated proctors would need to be supplied, alternative means for transmitting the assessment would need to be used (hand courier, etc.), a facility manager cleared for access to the materials would need to be present to provide any needed support in using the facility, procedures to remove materials from the workstations and local server would need to be reviewed with considerable scrutiny. Separate procedures may need to be evolved for storing that data on centralized servers and for security of the assessment items. Any data analysis performed for validation of test items for such assessment would require separate handling. This is an issue that would require further consideration in depth. However, it is immediately clear that distance delivery of such items would be unwise.

- Assessing the capacity for on-site database and application servers in support of testing software when assessments:

  a) May have needs for timed delivery or which collect response time data measured in fine increments,
  b) Security sensitive content which makes distance delivery too vulnerable to consider,
  c) Or, assessments may have unique content that would stress bandwidth capabilities for distance transmission.

- Connectivity and bandwidth
- Local technical support services
- On-site servers for assessments that may not be appropriate for distance delivery.
- Proximity of testing centers to the potential participants
- The availability of appropriate measures within the testing center to frustrate efforts at test compromise and cheating

## SCORM Requirement

One requirement imposed by DLS is that the application (training or testing) be compliant with the sharable content object reference model (SCORM). SCORM establishes a standardized way of producing courseware so that it is playable. Still evolving in its technical specifications, it allows for the creation of a huge repository of courses and training-related products that will greatly facilitate access for all of the services.

## Concept of Operations for DCAP

Soldiers to be tested will be identified through ARI's usual research support request (RSR) process and/or through negotiations with Army commands and National Guard units. Once tasked, units will need to provide identifying information about the Soldiers who will be tested. Participating Soldiers will be instructed to contact the local DTF manager and schedule an appointment for registration in the DLS system. Registration takes about 20 minutes per Soldier and requires input from both the Soldier and the DTF manager. After the registration is successfully completed, the Soldier will schedule a date and time to take the test that is within the timeframe established for the pilot test. For purposes of the DCAP, ARI and HumRRO personnel will likely serve as proctors for the DTFs that will participate in the pilot test.

Local units will also provide the names and Social Security Numbers of selected Soldiers to ARI who will load the DCAP testing requirement into the training management system. This step will only allow registration for DCAP by authorized Soldiers.

The assessment will reside on a secure server. In addition, access to the test will require authentication through the DLS central server. The DLS server will generate a unique ID that it will pass to the test administration server. This ID will allow the connecting of test data to identification information from within Army systems. Thus, the server interaction will provide security to the test content and achieve Privacy Act requirements.

## Full-Scale Implementation

There are a number of issues to be addressed before implementation of Army-wide promotion testing. These include:

- DTF capacity – Currently, DTFs have available capacity to support small to moderate computer-based testing programs. However, universal promotion testing could easily swamp the available seat hours at existing sites. In addition, the planned expansion of DTFs may have to be accelerated to serve all Soldiers. It is possible that Soldier assessment could become the predominant activity at DTFs, and with this, the need to provide direct funding support to DLS.

- Assessment Management systems – In the short run, assessments are being managed like training packages within overall training management structure. That is, Soldiers are identified to receive require training, the system provides appropriate notices to Soldiers, and tracks progress toward completion. It is not clear that this approach could work in a large-scale implementation.

- Internet versus local server assessment administration – Internet administration is great for text-based (i.e., low bandwidth demand) assessments. However, as the graphics, audio, and video content in the assessment increases, internet administration may not be feasible. The alternative is to move the assessment to a local server. This mode of administration would require the development of new systems within the DTF operation.

- Proctoring requirement – It will be necessary for the Department of the Army to arrange to have trained proctors to monitor all operational testing that takes place at the DTFs. The DTFs do not currently have sufficient staff to meet this requirement.

All of these challenges are formidable, but the Department of the Army has the foundation for a large-scale assessment process already in-place.

## Web-Basing Assessments

### Introduction

While the beneficial aspects of web-based assessments are apparent, it is also clear that the deployment of such technology will present areas in which planning, efficient resource management, and thoughtful consideration will be required. Fortunately, the Army already possesses many of the building blocks, both in facilities and personnel, which will be crucial to the success of this effort. These assets may be leveraged in a manner that will prove beneficial to the overall goal of making web-based assessments more universally available to all Army specialties. However, there will likely be areas where a need to grow new support structures will emerge. An effort will be made here to describe the early impressions of what this effort may entail and to begin a discussion of the concerns that may need to be addressed as matters progress. As the understanding of the work grows, we anticipate that new concerns may emerge that will be important to address as part of this discussion.

### Key Considerations

Delivery of web-based assessments requires certain server-side components in order to function. Generally, the assessment itself and any associated software platforms reside on a server that is configured for web delivery of content. Additionally, these applications often require a database application which houses assessment materials and/or student responses.

Functionality is significantly enhanced when the application software resides on a separate server from the database application. This permits a greater range of resources for processing the transactional demands of the database application. Many software vendors will go so far as to stipulate that this is a required practice. Regardless, it is an industry wide best practice.

When discussing server facilities one must consider the hardware components (the servers themselves, mechanisms for failover redundancy and backup equipment), the software components (server operating systems, application software and database software), personnel (staffing to ensure that the equipment is properly maintained, and staffing to ensure that the applications are updated and functional), connectivity (provision of bandwidth that accommodates the distribution and receipt of content), licensure (for applications utilization), and physical location (adequate room for equipment, proper climate control, fire suppression, security, etc.).

Some specific areas of concern include:

- Sufficient server resources to deliver the assessments
- Sufficient server resources to house database resources in support of the assessments
- Separate housing of assessment server software and database servers.
- Local technical support services
- Sufficient skilled personnel to facilitate maintenance of the applications, resolve potential conflicts, schedule updates, facilitate successful data delivery to appropriate personnel systems when appropriate, and provide assistance to the field.
- Connectivity and bandwidth resources appropriate to the task.
  - Storage and transmittal of assessment materials

Software requirements in support of test delivery include:

- Perception Server
- Questionmark Secure Browser
- MS Windows 2000 Server
- MS SQL Server

*Server Facilities/Software Resources*

The current testing model utilizes Perception as a platform for assessment item creation, delivery, and as the mechanism for storing data collected from the assessments. While this is not the only possible choice for use by operational assessments, it is optimal in terms of consistency, operability, and in its capacity to support on-going development as new assessments are added and evaluated. As hardware and software concerns relate directly to one another, they will be grouped for discussion

The Perception platform can be supported on the Army's servers, or the Army has the option of purchasing hosting with the software vendor directly should they wish to forgo the responsibility of housing the application in-house. Either choice can be seen as advantageous to the Army and as supportive of continued development and deployment of assessments.

Each choice also has its limitations. An Army hosted solution requires infrastructure expense and personnel requirements in manpower, hardware, and software. It also requires some convoluted work-arounds between the existing personnel system, training managers and the assessment developers, and researchers in an effort to protect privacy data (such as social security numbers) when identifying participants and associating scoring data with participant IDs.

During the operational phase of the assessment, this will continue to be a concern for test developers and researchers who are then trying to harvest information from the assessments to use in item validation and other research (e.g., criterion-related validation projects). Such data must be returned from the Army system, stripped of identifying SSNs. The current Army system is not configured to support the test development/ research function conveniently. The system will require a routine way for feeding item-level data back to test developers and for incorporation into a computerized item bank and for updating assessments. There will also be a need to accommodate periodic requests from researchers in a way that is timely but does overburden Army personnel who are overworked or otherwise tasked.

Turning hosting over to the software vendor for delivery also has some downsides. There will be a cost for delivery of this service. This is an area we have not really begun to explore. Also, a third party is introduced into the equation, which will add a layer of complexity to coordination between the Army, assessment developers and ARI researchers. And, the concerns for data housing are exacerbated, as it seems much less likely that the Army will want its personnel information, raw data, or other sensitive content shared with a third party.

Hosting of the application for long term, operational deployment by HumRRO or another contractor does not seem appropriate at this time. However, it would have the advantage of allowing direct access to researchers and test developers, and increases flexibility in making alterations to or deploying additional assessments.

This said, the extent of resources needed for web delivery during the operational phase remains unclear. Fortunately, the phased nature of the operational stage will make it possible to grow the infrastructure sensitively as assessments become available.

At a minimum, some recommended requirements for delivery infrastructure are:

- Appropriate licenses for the Perception Server component.
- Appropriate licenses for the secure browser component.
- An application server platform running Microsoft Windows Server 2000. This cannot be a UNIX or LINUX server. This houses the Perception Server component and assessments.
- A database server running either Microsoft Windows Server 2000 in support of a MS SQL database, or a UNIX server running Oracle. (It is important that this be a requirement coordinated on between the researchers and the system administrators (Army) hosting the assessments. Otherwise, additional complexity will be encountered if an assessment is developed with an SQL backend and the deploying system will be using an Oracle database.)
- A scheduling mechanism that avoids assessments being scheduled with identical start or end times for more than 50 users at a time. (This being the manufacturers' guidelines for minimizing the chance of overloading the servers. While hundreds or even thousands of assessments may take place over the course of a day comfortably,

there are operational limits to the practicable number of synchronous assessments before significant and noticeable delays in delivery become apparent. This could have significant implications in the case of timed assessments. If this is unavoidable, options for load balancing with more than one application/database server team must be explored.)

- Co-locating the Perception Server application on a server housing other applications with any significant usage is highly discouraged.
- In the long term, an additional server may be required to house the Enterprise Manager function of the software that would allow generation of reports and data usable by research personnel.
- Linkage between the database server and the application server(s) must be at a minimum speed of 100Mbps or greater.
- Servers running Microsoft Server 2000 should not be dual or multi-processor units or with FAT32 file systems.
- System memory minimums are at 512MB. This is a vast under-representation of the requirement. 2-4 GBs is suggested.
- Further exploration is required to fully understand the bandwidth loads associated with this application. Each assessment may need to be considered separately due to differences in the numbers of assessment blocks, included graphics, calls to helper applications (such as Java or MathML, Flash, etc.), and so on. This may effect recommendations for how many centers/participants may take a selected assessment simultaneously, even assuming T1 lines are in evidence. For a typical multiple-choice assessment with few inclusions, this should not be a difficulty for the current level of technology displayed by the DTFs; however, the inclusion of specialized content or simulations may significantly impact capacity.
- It is also unclear how much storage will be required as these assessments become operational. Not only is storage an issue for active test items, assessments and scoring; but it may be a concern for archival purposes. *No discussion has yet been held about the desirability of retaining data for research purposes or for the purposes of test challenges by a participant.* This is an important issue that needs to be resolved.
- If it is seen as necessary for local facilities to deliver content to avoid long distance connectivity issues or for other purposes, a minimalist version of this system will be required at each testing facility ***(the location of at least one application and database server at each location where assessments containing specialized content are scheduled for delivery)***. Thought needs to be given to all the possible ramifications of that requirement and the staffing needs that it may generate.

### *Administrative Support*

### Key Considerations

To support long-term assessment, the DTFs will need to staffed with personnel who can support the technical aspects of test delivery and who can proctor the exams. Specifically, the following will be required:

- Personnel to support scheduling, problem resolution, and liaison services between the assessment centers, centralized assessment storage and delivery centers, existing personnel tracking centers, and customers/users at the unit level.
- Augmentation or additional training of assessment center staff to accommodate the demands of the assessment delivery program. To focus on procedures, familiarization with the delivery software, how to assist the user/what assistance is appropriate, security concerns, the role in ensuring any local assessment materials are downloaded and prepared for use, responsibilities they may have in seeing that the completed assessment information is packaged and the sent (physically or electronically), interfacing with proctors, etc.
- Trained proctors to ensure that examinees take the assessments according to administrative procedures and rules.

As the Perform M21 development effort moves into the field testing and eventually operational delivery stages, a number of on-going requirements will be generated in administering and managing this effort. It seems likely that this will result in the augmentation of staff at any centralized facilities for distance learning content dissemination, for the distance training facilities themselves, and in any section which actively manages the interaction between learning managements systems and personnel data.

*Summary and Conclusions*

The utilization of web-based assessments for distance and localized delivery to Army personnel offers some specialized concerns for test developers, Army researchers, and resource planners, yet carries with it a great expectancy for success through the employment of existing DTF resources and carefully considered application of additional resources.

How new assessments are added to the body of assessments that are operational, the procedures for implementing changes or discontinuing invalidated questions once assessments have reached the field, and who would be involved in implementing these changes are all issues that have yet to be clarified. Procedures for taking data captures in support of research efforts once the assessments become operational are also in need of consideration.

In order for this effort to proceed cooperatively and positively, these are areas where attention is needed to forge a durable understanding and working relationship beyond the somewhat temporary and makeshift accommodations that are employed in the experimental (DCAP) phase. While Army personnel and researchers are both making their utmost effort to be accommodating in the short term, these temporary solutions will quickly erode under the strain of the operational testing as individuals who have extended themselves so generously must feel constrained to withdraw their support due to other pressing workload requirements. Attention needs to be given to support for these taskings at the earliest opportunity.

A number of decision paths need to be explored in the near term, the results of which will inform current project activities. One such issue would be the suitability of hosting assessments with the software vendor that developed the Perception program currently in use by the bulk of the project's assessment designers. Another might be organizing a methodology for rating individual assessments according to their technical resource requirements for bandwidth, local v. distance delivery, sensitive content, etc. Finally, it seems critical that a mutually agreeable protocol for coordinating new content additions and researchers' activities with Army stakeholders be developed. Despite the technical nature of many of the challenges posed by the implementation phase of this project, a protocol for on-going collaboration may turn out to be the most important resource that can be developed in order to ensure the long term technical success of the project; and, unlike many of the technical requirements, this protocol and the relationships it supports is not amenable to monetary or hardware driven solutions.

The existing Department of the Army infrastructure provides a strong foundation for the deployment of the project, but will experience some strain as the DTFs move from largely under-utilized facilities to facilities that must accommodate the volume of assessments that are currently envisioned. It would probably be sensible to begin planning a phase in of the resources needed to support the assessments now, to avoid publication of assessments (and stakeholder demand) which outstrips the preparation for their delivery, thus decreasing the likelihood of resorting to costly, temporary fixes.

Training needs and projected personnel requirements might also be mapped in a similar manner. It may prove beneficial for a number of the DTF managers to hold a discussion about the impact that this process is likely to have with a goal of generating a list of resource requirements and a timeline, which appears sensible for their implementation. In this way, assessment developers and Army stakeholders have a greater opportunity to work cooperatively with fewer frustrations.

While further work remains in obtaining a more thoroughgoing understanding of existing Army processes and organizational structures as they relate to the web-based delivery process, initial impressions suggest an optimistic assessment of the project's capacity to employ web-based methodologies in support of PerformM21.